Which topics are visualized by science maps? A topic-driven clustering effectiveness analysis

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

[1]Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, the Netherlands
[2]Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, the Netherlands

Abstract

In this paper, we analyze the effectiveness of science maps based on citation similarity for clustering biomedical topics, in particular MeSH terms. Science maps group documents by topic to provide an academic literature overview. Because documents have multiple topics, some topics will correspond to the document clusters, while other topics might not be represented by a cluster. To identify these topics, we analyze the extent to which document clusters are formed about topics and use this as a signal to evaluate the topical clustering effectiveness. We found that the best clustered topics are related to diseases, organisms, anatomy, and techniques and equipment for diagnostics and therapy, while the worst are related to geographical entities, information sciences, natural science fields, and health care and occupations. Our findings indicate that science maps are better at clustering some topics than others, which is relevant for science maps users. The analysis method that we developed is also a contribution to topic-driven evaluation of science maps.

Keywords

Science maps, Clustering, MeSH terms, Topical analysis

1. Introduction

Science maps [1] are visualizations that provide an overview of the content of academic documents collections. The goal of science mapping is to find meaningful structures in the data sets of scientific publications, which can then be used for literature analysis or information retrieval [2, 3]. Some of the uses of science maps are field delimitation [4], research policy [5], and enhanced document browsing [6]. A well established practice to create science maps is to cluster similar academic documents together, and then to summarize the content of these documents. In other words, the map is a set of clusters that emerged from document similarity, and the topic of the clusters is inferred from the documents it contain.

When using science maps, it is important to remember that academic documents can have more than a single topic (e.g. a document about *lung cancer* is both about *lungs* and *cancer*), but in a science map they typically can be assigned to only one cluster with a single topic. Losing information when creating overviews is unavoidable, but it does raise the question of which of the topics (aspects) of the documents the clustering will be based on. This is not an idle question,

Joint Proceedings of BIR 2024: 14th International Workshop on Bibliometric-enhanced Information Retrieval and IR4U2 2024: 1st Workshop on Information Retrieval for Understudied Users
j.p.bascur.cifuentes@cwts.leidenuniv.nl (J. P. Bascur); s.verberne@liacs.leidenuniv.nl (S. Verberne)
0000-0002-4077-1024 (J. P. Bascur); 0000-0002-9609-9505 (S. Verberne); 0000-0001-8448-4521 (N. J. v. Eck); 0000-0001-8249-1752 (L. Waltman)

as there is an increasing disagreement between expert-identified topics and clusters-identified topics [7], indicating that the expert-identified topics are poorly represented. More specifically, experts find that their identified topic documents are placed on clusters where they are minority, instead of forming clusters where they are majority.

In the current work, we investigate clustering for biomedical topics – in particular MeSH terms, on clustering solutions based on citation similarity. We aim to find out which of the MeSH terms form clusters where they are majority, a phenomenon that we will refer as *clustering effectiveness*. Our approach is to group the biomedical topics into topic categories, and measure the clustering effectiveness of each topic, so to create a ranking of topic categories.

Our main research question is: Which topic categories have the highest and lowest clustering effectiveness in science maps? Our contributions are twofold:

- We develop a methodology to compare the representation of different topics categories in a clustering solution.
- We find that some topics categories that are better represented in a science map, and which topics categories these are.

2. Related work

Klavans and Boyack [8] analyzed how the areas of science relate to each other by aggregating several maps of science that were created using different methods. Ruiz-Castillo and Waltman [9] analyzed a method to normalize the citation count of documents according to the citation count of their neighboring documents. This normalization method is used by the Leiden Ranking to evaluate universities scientific production [10]. Science mapping has also been used for information retrieval purposes, by grouping the documents in hierarchical levels to facilitate the navigation of the map at different granularity [11, 12, 13, 14, 6]. Creating new groupings of documents allows the user to increase the topic diversity of the retrieved documents [15].

The most common method to evaluate the quality of a map of science is to ask experts from the scientific fields of the documents if they think that the map of science reflects their knowledge about the fields. The utility of this evaluation method is limited because it usually gives an inconclusive result: The experts tend to agree with most of the science map but identify caveats about certain details [16]. Additionally, there are several problematic issues intrinsic to the expert evaluation method: The evaluation criteria may change between experts; seeing the map may affect the expert own interpretation of the fields; the expert may be biased towards the fields of their interest; and the expert may have limited competence in some (sub)fields [16].

An alternative method to evaluate the quality of a map of science is to consider the intrinsic properties of the network, which does not require additional metadata about the documents. Commonly used intrinsic properties are specific desirable characteristics: 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 [17]. Another approach to evaluate a map by its intrinsic properties is to assume that there exists an ideal map that can be intuitively recognized by users and that any similarity metric that one can calculate between documents is a proxy of the similarity metric that would generate the ideal map [18]. Ahlgren et al. [19]

used this method to measure the accuracy of several similarity metrics using a MeSH terms similarity network for evaluation.

The third approach to evaluate the quality of a map of science is to define a ground truth made of documents that correspond to a given field, and evaluate the overlap between the clustering solution and the ground truth: either to which extent all the documents of each field are contained in a single cluster [7, 20], or to which extent each cluster contains only documents of a single field [21, 22, 23]. Some works obtain the ground truth from the references of review articles [24, 13], but most works obtain the ground truth using expert knowledge. To our knowledge, MeSH terms have not been used as ground truths, but instead have been used as one of the tools to support expert knowledge ground truths.

Evaluations that use expert knowledge ground truths have recently question the quality of maps of science by challenging their capacity to identify fields of science [23, 7, 20, 22]. One explanation for this negative result is that it happens because a single document can belong to several fields but it can only belong to a single cluster [7, 22], although some maps allow documents to belong to multiple clusters [25, 26]. Another explanation is that the use of different clustering algorithms can have a significant influence on the quality of the maps, and it is impossible to know beforehand which clustering algorithm will give the best result for a given map [27, 21]. In this paper, we investigate the clustering effectiveness of MeSH terms as a start to close this gap.

3. Methods

This section has the following structure: In subsection 3.1, we define how we select the documents, clusters, MeSH tree branches, and MeSH terms that we use in our analysis. In subsection 3.2, we explain how we measure the clustering effectiveness of the topics.

3.1. Data selection

Documents The collection of documents that we use in our work comes from the work by Ahlgren et al. [19]. 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., who clustered the above-mentioned documents using different similarity metrics and granularities, and the Leiden algorithm [28] for clustering, where the parameter Resolution controls the granularity of the clustering solution (a higher Resolution value generates smaller clusters). We selected the clustering solutions that Ahlgren et al. generated with the metric *Extended direct citation* (EDC), which is calculated using the direct citation between documents plus the citations to documents outside the document collection [18]. We selected clustering solutions based on this metric because it has a good performance in the work by Ahlgren et al. From the EDC based solutions, we selected the clustering solutions of three Resolution values: $2 * 10^{-6}$, $2 * 10^{-5}$ and $2 * 10^{-4}$, because the first and second values yield cluster sizes similar to those in the CWTS algorithmic mapping of science [11] (which is used in the Leiden Ranking [10]), while the third value enables us to evaluate clusters of smaller size.

Table 1
Statistics of the clustering solutions.

Resolution	N. clusters	Median cluster size
$2 * 10^{-6}$	297	7,615
$2 * 10^{-5}$	2,469	878
$2 * 10^{-4}$	21,659	88

We cleaned the clustering solutions by removing the clusters with a size of 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.

Topics Our topics are the Medical Subject Headings (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 [29]. We obtained the MeSH terms annotated for each document in our document collection, plus the metadata of the MeSH terms, from the in-house CWTS database of PubMed and MeSH (version from 2023).

We expanded the MeSH terms per document using the functionality *exploding MeSH terms* of Pubmed, where a MeSH terms query retrieves not only the documents annotated with the query MeSH terms but also the documents annotated with a descendant of the MeSH terms in the MeSH terms tree. The result of this search is equivalent to adding to the documents the MeSH term parents of the MeSH terms annotated in the documents and then performing the query. We replicated the latter process by annotating the parents of the MeSH terms to the documents. A single MeSH term can belong to more than one MeSH term tree branch (whose importance is explained below), so to avoid documents connecting to additional branches through the additional parent MeSH terms, we included in the parent MeSH terms the branches of the annotated MeSH terms they came from.

We also removed the annotated parents that where too redundant with their children, that is, were present in about the same documents. We did this by creating groups of MeSH terms with a Jaccard similarity of at least 0.9 with each other, where the similarity is calculated as the number of documents where both MeSH terms are present, and then removed all but the MeSH term with the least documents from these group. After that, we made three groups of MeSH terms according to number of MeSH term documents they have, which we will refer as Size bins: 501-1,000, 2,001-4,000, and 8,001-16,000. We made Size bins because the number of MeSH term documents affect the evaluation, as explained in the next subsection, and we chose these three Size bins because they contain a high number of MeSH terms, while also being spread apart from each other in terms of number of MeSH term documents.

Topic categories Our topic categories are the 16 nodes at the first level of the MeSH hierarchical tree of topics [29], also known as the branches of the MeSH tree. The MeSH terms *Male* and *Female* are not part of any branch, so we will ignore them in our analysis. We use branches because they group the MeSH terms in epistemological categories (e.g. organisms), which are

Table 2
Branches considered in the analysis. The code is the MeSH terms tree code for the branch. The color of the code is the same as in Table 3. The Size bin columns indicate the number of MeSH terms that belong to each Size bin, for each branch.

Code	Branch name	Number of MeSH terms per Size Bin		
		501 - 1,000	2,001 - 4,000	8,001 - 16,000
A	Anatomy	209	161	76
B	Organisms	247	98	44
C	Diseases	472	272	114
D	Chemicals and Drugs	1,033	568	264
E	Analytical, Diagnostic and Therapeutic Techniques, and Equipment	324	253	150
F	Psychiatry and Psychology	109	95	38
G	Phenomena and Processes	264	221	143
H	Disciplines and Occupations	50	31	15
I	Anthropology, Education, Sociology, and Social Phenomena	57	40	24
J	Technology, Industry, and Agriculture	76	68	26
L	Information Science	31	28	18
M	Named Groups	21	20	14
N	Health Care	182	134	87
Z	Geographicals	51	36	21

the categories sometimes used for topical analysis of clusters [7, 30]. We remove the branches with fewer than 10 MeSH terms per bin. We ended up with the 14 branches shown in Table 2.

3.2. Clustering effectiveness

Selection of clusters In Section 1 we loosely defined clustering effectiveness as the extent that the documents of a given MeSH term form clusters where they are majority, which can be interpreted as the extent to which the MeSH term documents create clusters about the academic topic. To measure this, we first identify the clusters in a clustering solution that contain the MeSH term documents. Then, we sort these clusters by number of MeSH term documents, from higher to lower, and sequentially select them until we cover half of the MeSH term documents. We only use the selected clusters in our evaluation because we want our analysis to be resistant to MeSH term documents with an outlier behaviour, which could be caused by poorly annotated MeSH terms or by very small clusters. Our cluster selection criterion is inspired by cluster quality metrics of Yuan, Zobel and Ling [31], and we expect that this criterion reflects the kind of clusters a user is likely to select while browsing documents in a science map.

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

1. 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:

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

where 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.

2. Inverse count of clusters (ICC): ICC represents the extent to which the MeSH term are contained in a small number of clusters. ICC is defined as one divided by the number of selected clusters. The higher ICC, the more effective the clustering. ICC is bounded between zero and one.

We use two metrics instead of only one so we can take the influence of Size bin and Resolution into account, as they affect each metric in opposite ways: Higher Size bin and Resolution increase Purity because they increase the probability of a given cluster to have a high concentration of documents, while they decrease ICC because they increases the number of clusters necessary to cover the MeSH term documents.

The Purity or ICC of a branch is the median metric of their MeSH terms for a given Resolution and Size bin. We intend to answer our research question by ranking the branches according to Purity or ICC, for each Resolution and Size bin combination.

4. Results

Table 3 shows the branches ranked by Purity for each Resolution and Size bin combination. This table also includes the Purity and ICC values of each branch for each combination. We did not show a table ranked by ICC because it is similar to Purity rankings. In this table we observe:

- The Purity and ICC of any given Resolution and Size Bin combination tend to rank the branches in the same order.
- The position of the branches tend to be similar among the Resolution and Size Bin combinations.
- On the top four positions, the following branches appear more than half of the time: C (Diseases), B (Organisms), E (Analytical, Diagnostic and Therapeutic Techniques, and Equipment) and A (Anatomy).
- On the bottom four positions, the following branches appear more than half of the time: Z (Geographicals), L (Information Science), H (Disciplines and Occupations) and N (Health Care).
- C and Z are always the highest and lowest branches, and their Purity and ICC are significantly higher and lower than the closest branch.

5. Discussion

In this section we discuss what we learn from the results from highest and lowest clustering effectiveness, and the strengths and weaknesses of our work.

Table 3
Purity ranking of branches for different Resolution and Size bin values.

Resolution	Size Bin	Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$2*10^{-6}$	501 - 1,000	Branch	C	A	E	B	M	F	I	G	H	J	D	N	L	Z
		Purity	0.027	0.012	0.010	0.009	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.006	0.003	0.001
		ICC	1.000	0.500	0.500	0.500	0.500	0.500	0.500	0.333	0.333	0.333	0.333	0.333	0.200	0.083
	2,001 - 4,000	Branch	C	B	E	A	F	I	G	J	N	M	D	H	L	Z
		Purity	0.082	0.036	0.031	0.027	0.022	0.021	0.020	0.020	0.017	0.016	0.016	0.013	0.011	0.003
		ICC	0.500	0.500	0.333	0.333	0.500	0.333	0.250	0.333	0.250	0.200	0.200	0.167	0.167	0.056
	8,001 - 16,000	Branch	C	B	J	A	E	F	I	G	L	N	D	M	H	Z
		Purity	0.187	0.082	0.072	0.062	0.053	0.053	0.052	0.051	0.034	0.032	0.032	0.029	0.021	0.012
		ICC	0.500	0.333	0.333	0.200	0.183	0.292	0.200	0.200	0.125	0.125	0.111	0.080	0.091	0.048
$2*10^{-5}$	501 - 1,000	Branch	C	B	A	E	F	J	M	I	D	G	H	N	L	Z
		Purity	0.142	0.059	0.043	0.036	0.033	0.032	0.032	0.028	0.028	0.025	0.018	0.017	0.012	0.003
		ICC	0.500	0.333	0.200	0.250	0.200	0.167	0.200	0.200	0.167	0.143	0.100	0.100	0.056	0.021
	2,001 - 4,000	Branch	C	B	E	A	J	F	M	I	G	D	N	L	H	Z
		Purity	0.262	0.160	0.074	0.073	0.062	0.061	0.054	0.049	0.049	0.045	0.042	0.031	0.031	0.005
		ICC	0.333	0.250	0.111	0.111	0.091	0.100	0.087	0.081	0.083	0.071	0.069	0.050	0.050	0.012
	8,001 - 16,000	Branch	C	B	J	F	A	E	I	G	D	L	N	M	H	Z
		Purity	0.337	0.171	0.134	0.116	0.111	0.105	0.094	0.090	0.075	0.070	0.065	0.040	0.035	0.016
		ICC	0.134	0.070	0.056	0.051	0.043	0.050	0.039	0.036	0.030	0.026	0.029	0.018	0.018	0.008
$2*10^{-4}$	501 - 1,000	Branch	C	B	A	E	J	D	F	I	M	G	N	H	L	Z
		Purity	0.337	0.193	0.106	0.091	0.082	0.080	0.080	0.076	0.073	0.068	0.042	0.040	0.033	0.013
		ICC	0.250	0.111	0.077	0.077	0.051	0.056	0.056	0.053	0.043	0.045	0.029	0.029	0.018	0.009
	2,001 - 4,000	Branch	C	B	E	A	M	J	F	I	D	G	N	H	L	Z
		Purity	0.410	0.344	0.146	0.132	0.128	0.124	0.118	0.106	0.103	0.097	0.082	0.070	0.068	0.018
		ICC	0.083	0.059	0.028	0.026	0.019	0.022	0.024	0.022	0.018	0.018	0.016	0.013	0.012	0.004
	8,001 - 16,000	Branch	C	B	J	F	A	E	I	G	D	N	L	M	H	Z
		Purity	0.519	0.313	0.243	0.191	0.186	0.173	0.148	0.148	0.145	0.121	0.119	0.084	0.074	0.031
		ICC	0.027	0.013	0.010	0.010	0.009	0.009	0.007	0.007	0.007	0.006	0.005	0.004	0.004	0.002

5.1. Which topic categories have the highest and lowest clustering effectiveness in science maps?

The similarity in the rankings between Purity and ICC suggest that the branches with higher Purity than others also have higher ICC, which is to be expected as both metrics reflect clustering effectiveness. The similarity in the rankings between different combinations of Resolution and Size bin value suggests that the clustering is systematically biased toward certain topic categories, with the strongest bias found at the top and bottom four positions of the ranking. This is a relevant result because the clusters might as well be indifferent to the topic category.

The top and bottom four branches are the most relevant information for a user of science maps because they inform which topics will be best identified by the clustering. We found a use case of sciences maps in the literature of which the findings agree with our results: Held and Velden [22] reported that, in a science map, documents from the field of invasive biology are spread among several clusters whose main topic is not this field but rather a given species. Invasive biology belongs to the branch H (Disciplines and Occupations), which is one of the bottom four branches, and species belongs to the branch B (Organisms), which is one of the top four branches. Therefore, we would have expected the science maps to be poor at showing the field of invasive biology, and to place the invasive biology documents in clusters about species.

It is specially interesting that H (Disciplines and Occupations) is among the bottom branches, because this branch contains the MeSH terms for natural science fields. The low clustering

effectiveness of this branch suggests a difference between how scientists define fields of science and how scientists cite each other. A solution could be to update how scientists define a scientific field by using science map clusters.

The positions of J (Technology, Industry, and Agriculture) and M (Named Groups) vary significantly in each ranking. The variance between different Resolution rankings is likely to be stochastic. The variance between different size bins could be due to the topic of their MeSH terms being very different between size bins. This is particularly likely for branch J because it has a broad scope of topics.

5.2. Strengths and weaknesses

Using a knowledge organization system of academic publications such as MeSH terms provides a massive number of topics, and very diverse compared to what a single expert could propose. Using a Biomedical knowledge organization system like MeSH terms instead of the ones from other fields, like the Mathematics Subject Classification, the ACM Computing Classification System, or the Physics Subject Headings, has the advantage that the Biomedical field produces the highest number of documents among the fields of science. On the other hand, a high number of Biomedical documents is detrimental to the creation of non-Biomedical clusters (the kind of problems that emerge in Internal Mapping vs External Mapping [22]), and we diminished this effect using a parameter for the Leiden algorithm (*Objective function: Constant Potts Model*) that is specifically intended to resist this effect. We also diminished the effect of mislabeled documents (e.g. the document *DOI: 10.1007/s12603-020-1457-6* is incorrectly labeled with the MeSH term *Alcohol Drinking*) by ignoring a share of the MeSH documents through the Coverage parameter. Exploding the MeSH terms while respecting their source branch made our labeling of documents more robust. Using the MeSH tree branches as topic categories allowed us to use a curated classification of topics. However, some topic categories might be absent from the MeSH tree (e.g. the topic that links diseases with their medicines) and some lower levels of the MeSH tree might work better as topic categories (e.g. the children of the the branch Disciplines and Occupations [H] are *Natural Science Disciplines* and *Health Occupations*, which might be more informative topic categories than the branch).

We only used the Leiden algorithm because it is the most commonly used by the science mapping community. However, similar works have used more than one algorithm: Held, Laudel and Gläser [7, 20] analyzed the topics in clusters created by the Leiden algorithm and the Infomap algorithm. Held [27] assessed the suitability of the Leiden, Louvian, OSLM and Infomap algorithms for creating topical clusters. Rossetti, Pappalardo and Rinzivillo [21] showed that different clustering algorithms (Louvain, Infohiermap, cFinder, Demon, iLCD and Ego-Network) have differential performance on different types of networks (DBLP co-authorship network, Amazon co-purchase network, YouTube users network, and LiveJournal users network). We only used one citation similarity metric (EDC), so it is impossible to say if our results are representative of all citation similarity metrics.

MeSH terms are oriented to Biomedical topics, which makes our findings difficult to generalize beyond the Biomedical domain. However, the same methodology can be used with knowledge organization systems oriented to other topics types, like computer science or astronomy.

Our work has the advantage that we calculate clustering effectiveness per MeSH term, while

other methods, like Waltman et al. [18], evaluate the clustering solution as a whole. Our method is also agnostic of the size relation between the MeSH terms and the clusters because we are trying to rank the clustering effectiveness of different topic categories instead of achieving optimal clustering effectiveness. Future work could ideally control for different MeSH term sizes by normalizing the Purity and ICC values instead of only comparing MeSH terms of similar sizes, but creating a normalizing function was beyond the scope of our work. The disadvantages of our method is that we lack a clustering effectiveness baseline to compare against, and that it is hard to compare our results with the literature because we are using a novel experimental setup.

6. Conclusion

In the current work we explored the extent to which the documents of given topic categories form clusters where they are majority, a phenomenon that we refer to as clustering effectiveness. The purpose of this is informing users of science maps what to expect from the maps. We found that some topic categories have consistently higher clustering effectiveness than others. Describing the topics with our own words, the highest topics categories are *diseases*, *organisms*, *anatomy*, and *techniques and equipment for diagnostics and therapy*, while the lowest are *geographical entities*, *information sciences*, *natural science fields*, and *health care and occupations*.

Our work has shown that science maps serve some topics better than others in a predictable way. Our analysis approach contributes to future topic-driven analysis of science maps. Future research should evaluate the clustering effectiveness of topic categories in science maps based on text similarity, which is another popular technique for making science maps after citation similarity.

Declaration of Generative AI

During the preparation of this work, the authors used ChatGPT in order to: grammar and spelling check, drafting content, improve writing style, citation management and formatting assistance. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] C. Chen, Science mapping: a systematic review of the literature, Journal of data and information science 2 (2017) 1–40. doi:10.1515/jdis-2017-0006.

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

[3] N. J. van Eck, Methodological advances in bibliometric mapping of science, EPS-2011-247-LIS, 2011.

[4] M. Zitt, Meso-level retrieval: Ir-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation, Scientometrics 102 (2015) 2223–2245. doi:10.1007/s11192-014-1482-5.
[5] R. Sullivan, S. Eckhouse, G. Lewison, Using bibliometrics to inform cancer research policy and spending, Monitoring financial flows for health research (2007) 67–78.
[6] J. P. Bascur, S. Verberne, N. J. van Eck, L. Waltman, Academic information retrieval using citation clusters: in-depth evaluation based on systematic reviews, Scientometrics (2023) 1–27. doi:10.1007/s11192-023-04681-x.
[7] M. Held, G. Laudel, J. Gläser, Challenges to the validity of topic reconstruction, Scientometrics 126 (2021) 4511–4536. doi:10.1007/s11192-021-03920-3.
[8] R. Klavans, K. W. Boyack, Toward a consensus map of science, Journal of the American Society for information science and technology 60 (2009) 455–476. doi:10.1002/asi.20991.
[9] J. Ruiz-Castillo, L. Waltman, Field-normalized citation impact indicators using algorithmically constructed classification systems of science, Journal of Informetrics 9 (2015) 102–117. doi:10.1016/j.joi.2014.11.010.
[10] CWTS, Leiden ranking fields, 2023. URL: https://www.leidenranking.com/information/fields, [Online; accessed 20-March-2023].
[11] L. Waltman, N. J. Van Eck, A new methodology for constructing a publication-level classification system of science, Journal of the American Society for Information Science and Technology 63 (2012) 2378–2392. doi:10.1002/asi.22748.
[12] P. Sjögårde, Improving overlay maps of science: Combining overview and detail, Quantitative Science Studies (2022) 1–40. doi:10.1162/qss_a_00216.
[13] P. Sjögårde, P. Ahlgren, Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics, Journal of Informetrics 12 (2018) 133–152. doi:10.1016/j.joi.2017.12.006.
[14] P. Sjögårde, P. Ahlgren, Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties, Quantitative Science Studies 1 (2020) 207–238. doi:10.1162/qss_a_00004.
[15] A. Nishikawa-Pacher, A typology of research discovery tools, Journal of Information Science (2021) 01655515211040654. doi:10.1177/01655515211040654.
[16] J. Gläser, Opening the black box of expert validation of bibliometric maps, in: Lockdown Bibliometrics: Papers not submitted to the STI Conference 2020 in Aarhus, 2020, pp. 27–36.
[17] L. Šubelj, N. J. Van Eck, L. Waltman, Clustering scientific publications based on citation relations: A systematic comparison of different methods, PloS one 11 (2016) e0154404. doi:10.1371/journal.pone.0154404.
[18] L. Waltman, K. W. Boyack, G. Colavizza, N. J. van Eck, A principled methodology for comparing relatedness measures for clustering publications, Quantitative Science Studies 1 (2020) 691–713. doi:10.1162/qss_a_00035.
[19] P. Ahlgren, Y. Chen, C. Colliander, N. J. van Eck, Enhancing direct citations: A comparison of relatedness measures for community detection in a large set of pubmed publications, Quantitative Science Studies 1 (2020) 714–729. doi:10.1162/qss_a_00027.
[20] M. Held, G. Laudel, J. Gläser, 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, 2020, p. 18.
[21] G. Rossetti, L. Pappalardo, S. Rinzivillo, 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, 2016, pp. 133–144. doi:10.1007/978-3-319-30569-1_10.
[22] M. Held, T. Velden, How to interpret algorithmically constructed topical structures of scientific fields? a case study of citation-based mappings of the research specialty of invasion biology, Quantitative Science Studies 3 (2022) 651–671. doi:10.1162/qss_a_00194.
[23] R. Haunschild, H. Schier, W. Marx, L. Bornmann, Algorithmically generated subject categories based on citation relations: An empirical micro study using papers on overall water splitting, Journal of Informetrics 12 (2018) 436–447. doi:10.1016/j.joi.2018.03.004.
[24] R. Klavans, K. W. Boyack, 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 (2017) 984–998. doi:10.1002/asi.23734.
[25] S. Xu, J. Liu, D. Zhai, X. An, Z. Wang, H. Pang, Overlapping thematic structures extraction with mixed-membership stochastic blockmodel, Scientometrics 117 (2018) 61–84. doi:10.1007/s11192-018-2841-4.
[26] F. Havemann, J. Gläser, M. Heinz, Memetic search for overlapping topics based on a local evaluation of link communities, Scientometrics 111 (2017) 1089–1118. doi:10.1007/s11192-017-2302-5.
[27] M. Held, Know thy tools! limits of popular algorithms used for topic reconstruction, Quantitative Science Studies (2022) 1–30. doi:10.1162/qss_a_00217.
[28] V. A. Traag, L. Waltman, N. J. Van Eck, From louvain to leiden: guaranteeing well-connected communities, Scientific reports 9 (2019) 5233. doi:10.1038/s41598-019-41695-z.
[29] N. I. of Health, Medical subject headings, Available at https://www.nlm.nih.gov/mesh/meshhome.html, ???? Accessed: 2023-09-07.
[30] C. Seitz, M. Schmidt, N. Schwichtenberg, T. Velden, 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), 2021.
[31] M. Yuan, J. Zobel, P. Lin, Measurement of clustering effectiveness for document collections, Information Retrieval Journal 25 (2022) 239–268. doi:10.1007/s10791-021-09401-8.