RESEARCH ARTICLE

A comparison of large-scale science models based on textual, direct citation and hybrid relatedness

Kevin W. Boyack[1] and Richard Klavans[2]

[1]SciTech Strategies, Inc., Albuquerque, NM (USA)
[2]SciTech Strategies, Inc., Wayne, PA (USA)

Keywords: accuracy, clustering, direct citation, hybrid similarity, relatedness measure, textual similarity

ABSTRACT

Recent large-scale bibliometric models have largely been based on direct citation, and several recent studies have explored augmenting direct citation with other citation-based or textual characteristics. In this study we compare clustering results from direct citation, extended direct citation, a textual relatedness measure, and several citation-text hybrid measures using a set of nine million documents. Three different accuracy measures are employed, one based on references in authoritative documents, one using textual relatedness, and the last using document pairs linked by grants. We find that a hybrid relatedness measure based equally on direct citation and PubMed-related article scores gives more accurate clusters (in the aggregate) than the other relatedness measures tested. We also show that the differences in cluster contents between the different models are even larger than the differences in accuracy, suggesting that the textual and citation logics are complementary. Finally, we show that for the hybrid measure based on direct citation and related article scores, the larger clusters are more oriented toward textual relatedness, while the smaller clusters are more oriented toward citation-based relatedness.

1. INTRODUCTION

With the increasing availability of large-scale bibliographic data and the increased capacity of algorithms to cluster these data, highly detailed science maps and models are becoming ever more common. Although most such models have been based on citation databases such as Web of Science (WoS; Sjögårde & Ahlgren, 2018; Waltman & van Eck, 2012) or Scopus (Klavans & Boyack, 2017a), open source databases such as PubMed are also candidates for such models. For instance, Boyack and Klavans (2018) recently clustered 23 million PubMed documents using openly available document–document relatedness data based on titles, abstracts, and Medical Subject Headings (MeSH) terms.

Recently, open source citation data covering a large fraction of PubMed have also become available (Hutchins, Baker, et al., 2019). This makes it possible to use citation and/or hybrid (text+citation) relatedness to create models of PubMed without linking PubMed data to citation data from one of the large citation databases. In this study we create and compare models of PubMed documents based on text, citation, and hybrid relatedness from the perspectives of relative accuracy and overlap. We find that although accuracy metrics are similar, there are substantial differences in cluster membership between models.

an open access journal

Citation: Boyack, K. W., & Klavans, R. (2020). A comparison of large-scale science models based on textual, direct citation and hybrid relatedness. *Quantitative Science Studies*, *1*(4), 1570–1585. https://doi.org/10.1162/qss_a_00085

DOI: https://doi.org/10.1162/qss_a_00085

Received: 18 May 2020
Accepted: 10 August 2020

Corresponding Author:
Kevin W. Boyack
kboyack@mapofscience.com

Handling Author:
Ludo Waltman

The MIT Press

2. BACKGROUND

2.1. Large-Scale Accuracy Studies

In addition to building on the large-scale modeling work mentioned above, this work also builds on increasing efforts to characterize the accuracy of models of science. Most accuracy studies have focused on relatedness measures rather than clustering methods. Few have attempted to compare the cluster-level results of such studies, with the work by Velden, Boyack, et al. (2017) as a notable exception.

Although those creating models of science have always sought to establish the validity of their models in some way, quantitative studies of accuracy are a more recent occurrence, particularly for large-scale models. The first such study using a very large literature data set was done by Boyack and colleagues using a set of 2.15 million PubMed documents (Boyack & Klavans, 2010; Boyack, Newman, et al., 2011). It compared text-based, citation-based, and hybrid relatedness measures where titles, abstracts, and MeSH terms were obtained from PubMed while references for each document were obtained from Scopus via matching database records. Among text-based approaches, the PubMed related article (RA) measure gave the best results when using grant–article linkages and textual coherence as two bases of comparison. The best citation-based and text-based approaches were found to have roughly similar accuracy. However, the hybrid measure outperformed all other relatedness measures, where the hybrid measure employed bibliographic coupling over a combination of references and words. Only words that occurred in at least four and not more than 500 documents were included in the calculation.

Years later, Klavans and Boyack (2017b) compared citation-based approaches using over 40 million documents from Scopus. They found that direct citation outperformed cocitation and bibliographic coupling using concentration of references in authoritative papers (those with at least 100 references) as the basis of comparison. However, their comparisons were normalized by numbers of clusters (i.e., were compared on graphs with numbers of clusters on the *x*-axis) and thus did not account for the cluster sizes.

In response to this shortcoming, Waltman, Boyack, et al. (2017) introduced a principled approach to comparing relatedness measures in which solutions are normalized using a granularity function that accounts for numbers of clusters and their sizes. This was done using a 10-year set of 272,935 publications from *Condensed Matter Physics* and comparing several different citation-based measures. In this study, bibliographic coupling, extended direct citation, and a combined direct/cocitation/bibliographic coupling all outperformed direct citation. Although this appeared to disagree with the results of Klavans and Boyack (2017b), in reality there was no disagreement, because the direct citation method of the latter was actually extended direct citation, which had not yet been so named. In addition to introducing granularity–accuracy (GA) plots, the principled approach suggests that the metric for comparison should be independent of all relatedness measures being compared to the extent possible.

Since the introduction of GA plots, all subsequent large-scale accuracy studies have adopted this principled approach to comparison. Sjögårde and Ahlgren explored topic-level (2018) and specialty-level (2020) models of over 30 million documents from WoS using direct citation. Rather than comparing relatedness measures, they assumed that direct citation was a reasonable approach and varied the clustering resolution—and thus the number of clusters—to identify the optimal resolutions and granularities for topics and specialties.

Boyack and Klavans (2018) created models of 23 million PubMed papers using the RA relatedness measure and compared accuracy statistics to those of the citation-based models they had reported previously (Klavans & Boyack, 2017b) using GA plots. Comparisons were

done using three separate metrics: concentration of references in authoritative papers, textual similarity, and grant-to-article linkages. The overall finding was that the textual RA scores produced models that were roughly as accurate as those created using extended direct citation. In addition, two RA models, one based on the top 12 neighbors and the other based on the top 40 RA neighbors, were compared and found to have similar accuracy. This suggests that edge filtering (restricting to the top *N* edges per paper) when using a textual similarity does not significantly degrade the accuracy of a model.

Waltman, Boyack, et al. (2020) recently updated their study to include data from two more disciplines (Cell Biology and Economics), confirming their earlier observations in these two disciplines. In addition, they found that edge filtering did not decrease the accuracy of their solutions, particularly for text-based relatedness measures.

The most recent study of hybrid relatedness measures is also the closest to our study in both intent and execution. Ahlgren, Chen, et al. (2020) compared 10 relatedness measures (six citation-based, one text-based, and three hybrids) using a set of 2.94 million documents from PubMed. References were obtained for each document by matching to WoS. Using a sophisticated weighting of MeSH terms as the basis of comparison, extended direct citation was found to perform the best, and slightly better than a direct citation + BM25 hybrid measure. Among simple measures, the BM25 text relatedness measure outperformed direct citation and cocitation, but was outperformed by bibliographic coupling. For the hybrid measure, three variations were used, with the best performance obtained with 33.3% as the text weight, followed closely by a 50% text weight. There was a significant drop-off when a 66.7% text weight was used.

Finally, in the interest of completeness we note that Haunschild, Schier, et al. (2018) evaluated the accuracy of clusters by comparing the contents of several microlevel clusters from Waltman and van Eck (2012) with the results of a keyword-based search, finding a significant lack of overlap between the sets. However, this is perhaps not surprising given that the microlevel clusters were based on citation relatedness while keyword-based query results are obviously textually oriented. Although this was not a large-scale study, its process could certainly be used at larger scale.

2.2. Hybrid Relatedness Measures

While some of the studies mentioned in the previous section have included hybrid relatedness measures, these have a relatively short history. Clustering of documents using citation-based relatedness measures (e.g., direct citation, cocitation, and bibliographic coupling) has been done for decades. The same is true for clustering using text-based relatedness measures. However, it was not until the mid-2000s that studies started to appear that clustered documents (e.g., Ahlgren & Colliander, 2009; Janssens, Quoc, et al., 2006) or journals (e.g., Janssens, Zhang, et al., 2009; Liu, Yu, et al., 2010) using hybrid relatedness measures. Studies published before 2010 typically used data sets much smaller (10,000 items or less) than those that are used today. These and other early studies are described in more detail in Boyack and Klavans (2010) and generally found that hybrid measures produced more accurate document clusters than nonhybrid measures.

Most early hybrid relatedness studies used linear combinations of text and citation relatedness scores or integrated approaches. Glänzel and Thijs (2011) did not agree with this approach and introduced an alternative based on the linear combination of angles (rather than raw relatedness values) from bibliographic coupling and tf-idf based on term frequencies. In subsequent work, Glänzel and colleagues suggested that the textual component should be weighted 16.7% (Thijs, Schiebel, & Glänzel, 2013) or 12.5% (Zhang, Glänzel, & Ye, 2016), with the bibliographic

coupling component taking the balance (i.e., most) of the weight. Later, Glänzel and Thijs (2017) found that multigram terms based on natural language processing (NLP) gave better results in this hybrid formulation than single words, and recommend that the textual component be weighted 75% when using NLP.

Other researchers also extended the work of Glänzel and colleagues. Meyer-Brötz, Schiebel, and Brecht (2017) generated over 100 relatedness measures using different weightings of text and citation components, first-order and second-order (vectorized) relatedness, edge filtering and numbers of clusters, finding that second-order relatedness with a 60% textual component gave the best results. They also found that the textual coherence of the resulting clusters increased with edge filtering and was maximized when only the top five or 10 nearest neighbors per paper were included in the clustering input. Yu, Wang, et al. (2017) used a combination of bibliographic coupling and cocitation, included typology features in their textual features, and accounted for reference age in their hybrid formulation, and suggested that the textual component should be weighted 45%. Although these studies were still rather small, with data sets of fewer than 10,000 papers, they nonetheless suggest ways to optimize hybrid relatedness that improve upon the performance of text-only or citation-only relatedness.

We have already mentioned the studies of Boyack and Klavans (2010) and Ahlgren et al. (2020), which employed hybrid measures on sets of millions of documents. In this study we expand upon those works and create and compare models of nine million PubMed documents based on text, citation, and hybrid relatedness using GA plots. We also address the question of how different the results stemming from different relatedness measures are in terms of overlap, a question that has until now been relatively unexplored.

3. METHODS

The general approach used in this study was as follows:

(a) A set of documents for the study was defined.
(b) Relatedness values between pairs of documents were calculated using seven different logics.
(c) Clustering was done using the Leiden algorithm (Traag, Waltman, & Van Eck, 2019) on each set of relatedness values, resulting in seven different models (sets of document clusters).
(d) Relative accuracy was then calculated for each model using three different bases of comparison (i.e., accuracy measures). Overlaps between models were also estimated.

3.1. Relatedness Measures

Our experiments make use of two different relatedness types: the direct citations available in the new NIH open citation collection, OCC (Hutchins et al., 2019) dated October 2019[1] and the RA (text relatedness) values calculated by the U.S. National Library of Medicine (NLM) based on the algorithm of Lin and Wilbur (2007)[2]. Seven different relatedness measures were calculated using these data. We did not include bibliographic coupling (BC) as one of the relatedness measures given its high computational cost (nearly 46 billion pairs before summing over pairs for our test set) and also, as others have found, that EDC performs as well (Waltman et al., 2020) or better (Ahlgren et al., 2020) than BC.

[1] https://doi.org/10.35092/yhjc.c.4586573
[2] https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Computation_of_Similar_Articl

Direct Citation (DC)

We define the DC relatedness between two documents *i* and *j* as

$$r_{ij}^{DC} = \text{avg}\left(c_{ij}, c_{ji}\right) \tag{1}$$

where c_{ij} = 1/*nref* if *i* cites *j* and is 0 if not, and *nref* is the number of references in document *i* within the OCC set.

Extended Direct Citation (EDC)

The EDC relatedness between documents *i* and *j* is calculated in the same manner as DC:

$$r_{ij}^{EDC} = \text{avg}\left(c_{ij}, c_{ji}\right) \tag{2}$$

where c_{ij} = 1/*nref* if *i* cites *j* and is 0 if not, and *nref* is the number of references in document *i* within the OCC set. The difference between DC and EDC is that for DC article *j* must be present in the test set, while for EDC article *j* can be outside the test set. By including cited articles outside the test set the number of direct citations considered is dramatically increased. DC and EDC relatedness values for all document pairs are bounded between 0 and 1. Note that we only included extended references that were cited at least twice. Also, our formulation of EDC is slightly different from that of Waltman et al. (2020). They modified the quality function used in the clustering algorithm to treat the extended articles differently than the main set articles, while we did not. Thus, we would expect our EDC results to be somewhat different from theirs.

Related Articles (RA)

NLM typically calculates RA scores (*S*) for each new document entering PubMed and keeps the top 100 scores. These are used to populate the "similar articles" that are listed on the PubMed webpage for each document and are also placed in a database for retrieval. Over a period of years, we have downloaded RA scores for all PubMed documents from the NLM using an Entrez query[3]. We use the top 20 RA scores for each document *i*, where

$$r_{ij}^{RA} = S_{ij} / \max\left(S_{ij}\right). \tag{3}$$

Normalizing by the maximum *S* value within the set bounds the relatedness values between 0 and 1. Also, because RA scores are symmetrical, in cases were document pairs *ij* and *ji* were both within the set, only the *ij* pair was included.

EDC+RA

Hybrid citation-text relatedness measures were created using EDC and RA relatedness as

$$r_{ij}^{EDC+RA} = \alpha r_{ij}^{EDC} + (1-\alpha) r_{ij}^{RA}, \tag{4}$$

where $0 < \alpha < 1$ weights the citation portion of the relatedness to achieve a desired mix such that $\Sigma\, r_{ij}^{EDC} = \Sigma\, r_{ij}^{RA}$.

Three variations were used in this study, with α = 0.8, 0.667, and 0.5 chosen because they compass a reasonable span over the values that have been found effective in previous studies (see section 2.2). Note, however, that the EDC-RA relatedness measure was not limited to the top 20 RA scores. Instead, RA scores for document pairs *ij* outside the top 20 were added where available. In addition, given that nearly all documents have some textual overlap with

[3] https://www.ncbi.nlm.nih.gov/books/NBK25499/, see cmd=neighbor_score

the documents they cite, RA scores that were not available were estimated as half of the minimum RA value for the corresponding citing document.

DC+RA

One hybrid citation-text relatedness measure was created using DC and RA relatedness as

$$r_{ij}^{DC+RA} = \alpha r_{ij}^{DC} + (1-\alpha) r_{ij}^{RA} \quad (5)$$

The parameter α is set such that $\alpha \Sigma r_{ij}^{DC} = (1-\alpha) \Sigma r_{ij}^{RA}$ to achieve a 50:50 weighting of citation and textual relatedness across the entire set of document pairs. As with the EDC-RA relatedness measures, additional RA scores outside for document pairs *ij* the top 20 were added where available.

3.2. Accuracy Measures

To compare the relative accuracies of the seven relatedness measures, we use three metrics that we have used previously (Boyack & Klavans, 2018): a concentration index based on references in authoritative papers, the fraction of RA similarity that ends up within clusters, and the fraction of document pairs referencing the same grant number that end up within clusters. Of these, the first metric is citation-related and is expected to favor the DC/EDC relatedness approaches, the second will obviously favor RA relatedness, and the third is independent of citation and text. Of these, the third metric satisfies the principled approach of Waltman et al. (2017) better than the other two metrics. However, the inherent biases of the first two metrics are known and, overall, the approach of using three metrics is balanced.

Herfindahl (concentration) Index

The first accuracy measure assumes that authoritative papers (those with at least 100 references) are written by authors who know their subject well, and that the references in those papers should be concentrated in relatively few clusters (per paper) in a more accurate classification system (Klavans & Boyack, 2017b). For the seven relatedness measures, we thus calculate an average Herfindahl index (H) using 29,470 papers published in 2017, each of which has at least 100 OCC references. The Herfindahl index for paper p in the model based on relatedness measure RM is

$$H_p^{RM} = \sum \left(s_k^{RM}\right)^2, \quad (6)$$

where s_k is the share (fraction) of references in cluster k. For example, if a paper has 100 references spread among three clusters with counts 70, 20, and 10, $H = 0.7^2 + 0.2^2 + 0.1^2 = 0.54$ for that paper in that model.

RA Fraction

This measure is very simple, in that we simply calculate the fraction of the RA values (limited to the top 20 per document) that is preserved within clusters as

$$F_{RA} = \sum r_{ij}^{RA} a_{ij} / \sum r_{ij}^{RA}, \quad (7)$$

where r_{ij}^{RA} is from Eq. (1) and $a_{ij} = 1$ if documents i and j are in the same cluster and 0 if not. This measure assumes that papers that have a strong textual link should end up in the same cluster, and thus that a higher value is associated with a more accurate cluster solution.

Grant-Based Fraction

Our final accuracy measure is one that is presumably independent of all the relatedness measures used in this study. Here we use articles linked to grants from the U.S. National Institutes of Health (NIH) and National Science Foundation (NSF), and from the UK Gateway to Research (GtR), which contains reports from multiple UK funding bodies. Data were retrieved from NIH ExPORTER link tables[4], the NSF project API[5], and the GtR website[6].

From the list of grant–article pairs, we created the full list of pairs of documents (limited to those published since 2000) that reference the same grant, resulting in a list of 248.6 million pairs. Although multiple funding bodies were used in this analysis, over 95% of the document pairs are associated with NIH grants.

The logic behind this accuracy measure is similar to the logic behind the other two measures. We assume that if two documents referenced the same grant, those two documents are likely to be similar and should appear in the same cluster. This logic is very reasonable for research project grants (such as NIH R01 grants) with a relatively narrow focus, but likely breaks down for large center grants that cover multiple topics. For this measure we calculate the fraction of document pairs that are preserved within clusters as

$$F_G = \sum G_{ij} a_{ij} / \sum G_{ij} \tag{8}$$

where G_{ij} is a document pair linked by a grant and $a_{ij} = 1$ if documents i and j are in the same cluster and 0 if not. Accuracy metrics are reported on granularity–accuracy (GA) plots, where granularity is defined as

$$G = N_{tot} / \sum N_k^2 \tag{9}$$

where N_{tot} is the total number of documents and N_k is the number of documents in cluster k.

4. DATA

For this study we chose to use PubMed documents, each represented by a unique PubMed identifier (PMID), from 2000–2018 that had at least 10 RA scores and at least 10 references to other PubMed documents in the OCC. Thus, each document had sufficient text and citation signal to not be dominated by one or the other. The set was not restricted by publication type.

PubMed contains 15,584,099 documents from 2000–2018 of which 11,691,296 (75.0%) have references in the OCC. The percentage of articles for which OCC has references is increasing over time, rising from 54.6% in 2000 to 86.9% in 2017. The gap between OCC coverage and Scopus coverage of references for the same articles (76.1% in 2000, 92.6% in 2017) is decreasing over time.

Upon limiting PubMed documents to those with at least 10 RA score and at least 10 references, and upon removal of articles from 42 chemistry and physics journals with little or no biomedical content, our test set consisted of 9,002,955 documents that contained 323,241,671 references to other PubMed documents, of which 185,517,801 were to PubMed documents within the test set. Our previous work with similar data showed that solutions based on the top 12 and top 40 RA scores per document had very similarity accuracy using the same three accuracy metrics (Boyack & Klavans, 2018). Thus, we chose to use the top 20 RA scores per document (152,379,249 in

[4] https://exporter.nih.gov/ExPORTER_Catalog.aspx?sid=0&index=5
[5] https://www.research.gov/common/webapi/awardapisearch-v1.htm
[6] https://gtr.ukri.org/search/project?term=*

Table 1. Characteristics of the seven different relatedness measures and associated models

	DC	EDC	RA	EDC+RA	EDC+RA	EDC+RA	DC+RA
α				0.80	0.667	0.50	0.50
# Test doc	8.99 M	9.00 M	9.00 M	9.00 M	9.00 M	9.00 M	9.00 M
# Cited doc		8.13 M		8.13 M	8.13 M	8.13 M	
# Pairs	185.5M	323.2M	152.4M	454.4M	454.4M	454.4M	316.7M
Resolution	7.813E-05	2.750E-05	2.188E-04	1.625E-04	1.063E-04	2.188E-05	3.875E-04
# Clust $\geq$ 50	21,881	16,315	20,128	15,551	15,826	15,377	19,348

total, deduplicated) as the base set of RA scores for this study. Characteristics of each of the seven relatedness measures are compared in Table 1.

Distributions of the DC, EDC, RA, and DC+RA relatedness measures are shown in Figure 1. DC and EDC relatedness values are very similar and range from 0 to 0.1 (with mean values of 0.02614 and 0.02772, respectively) as the test set was restricted to documents with at least 10 references. The RA distribution is shifted to the right with mean value 0.08667, while the DC+RA distribution is between the two, but with a mean value (0.07838) that is closer to the RA mean than the DC mean and an upper tail that extends beyond the tail of the RA measure. The DC+RA distribution is not exactly midway between the DC and RA distributions because the 50/50 hybrid is based on summed weights and the RA measure has fewer pairs than the DC measure.

Clustering was done using the Leiden algorithm for each input set, thus creating seven different models or cluster solutions. We desired roughly 20,000 clusters of at least 50 documents for each model to enable comparison; thus, resolutions for each model were set to target 20,000 clusters and are given in Table 1. Note that the EDC and EDC-RA models included 8.13 million cited

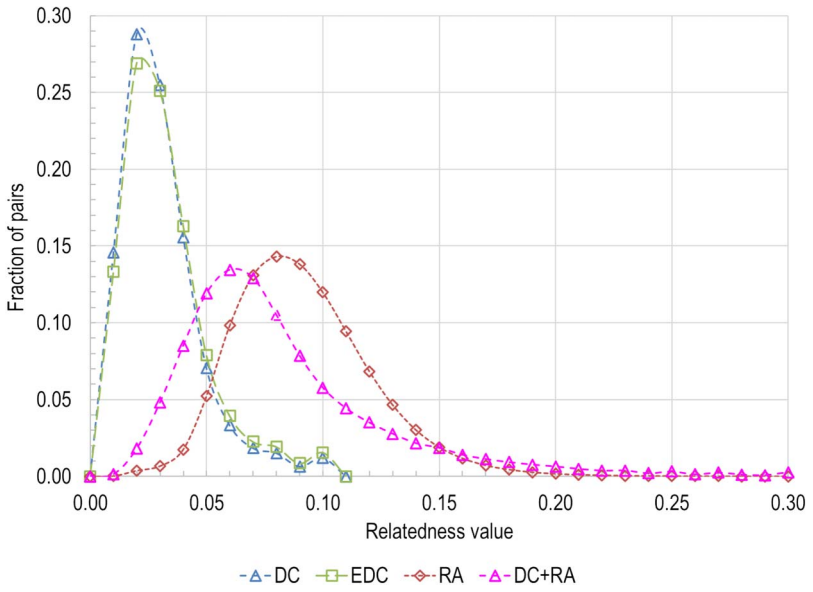

Figure 1. Relatedness value distributions.

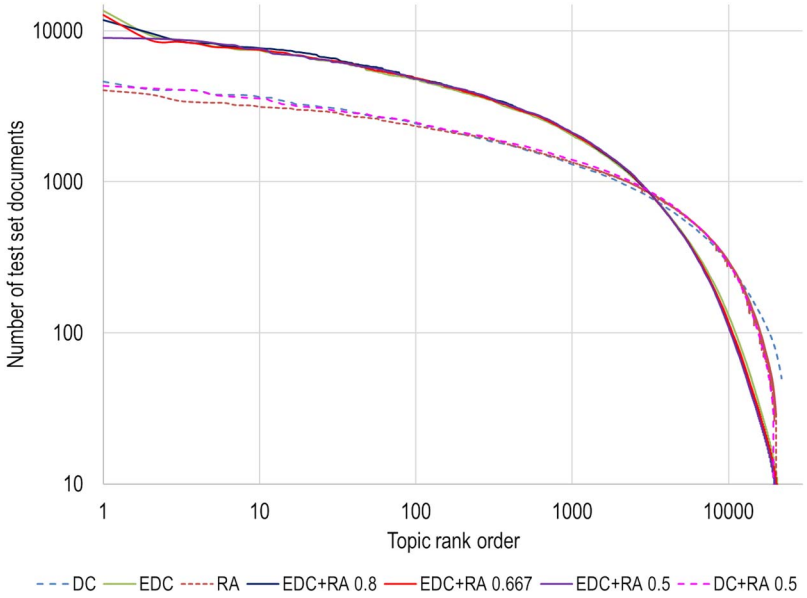

Figure 2. Cluster size distributions based on the 9 million papers in the test set. Extended papers from the EDC solutions are not included in the cluster sizes.

documents in addition to those in the test set. These were included in the initial clustering but were then removed before distributions and accuracy measures were calculated. These four solutions initially had over 20,000 clusters, but after removal of the cited (nontest set) documents the numbers of clusters with at least 50 members dropped by about 5,000.

Figure 2 shows the resulting cluster size distributions. For the DC, RA, and DC+RA models, the distribution is relatively flat with a largest cluster of over 4,000 documents. The four models that include EDC have distributions that are quite different—because the clustering included far more papers, the distributions are less flat with much larger clusters at the high end and a larger number of small clusters (counting test set documents only). When compared to DC, EDC may produce overly large clusters at the expense of smaller topics, and there may be too many smaller topics once the extended papers are removed. Based on our experience, a flatter cluster size distribution is a desired feature.

For each of the seven models, additional calculations were run to aggregate (in a hierarchical sense) the cluster solutions by factors of roughly 10 and 100, thus resulting in solutions of roughly 2,000 and 200 higher level clusters, respectively. These calculations were done to provide a range of resolution values so that curves could be compared on the GA plots rather than single points. We note that this manner of creating solutions over a range of granularities differs from that employed by Waltman et al. (2020) and Sjögårde and Ahlgren (2018, 2020). They run complete models for each granularity, while we aggregate hierarchically because we desire to have hierarchical solutions for practical analysis purposes.

5. RESULTS

5.1. Relative Accuracy

In this section the experimental results using three accuracy metrics are presented for the seven models along with a composite result. Figure 3 shows the average Herfindahl index over the

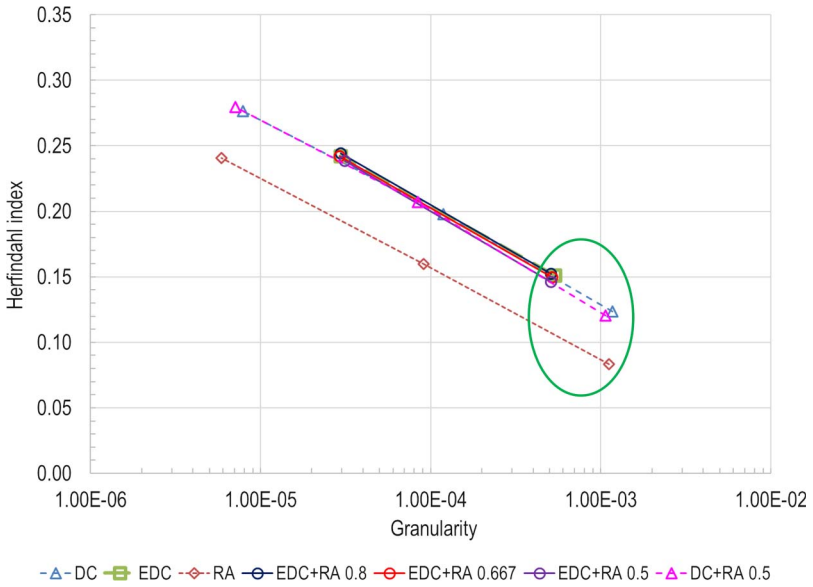

Figure 3. Herfindahl index values for the seven PubMed models.

baseline of 29,470 authoritative papers for each model at different granularities. Models at the resolution values noted in Table 1 are highlighted in a green circle. As expected, the Herfindahl index based on references substantially favors DC and EDC relatedness over RA relatedness. However, the hybrid relatedness approaches perform almost as well as the DC approaches. There is relatively little loss due to adding a textual component to the citation-based relatedness.

Figure 4 shows results for the fraction of RA scores preserved within clusters for each model. As expected, the RA metric favors the pure text approach, and the relative accuracy decreases using

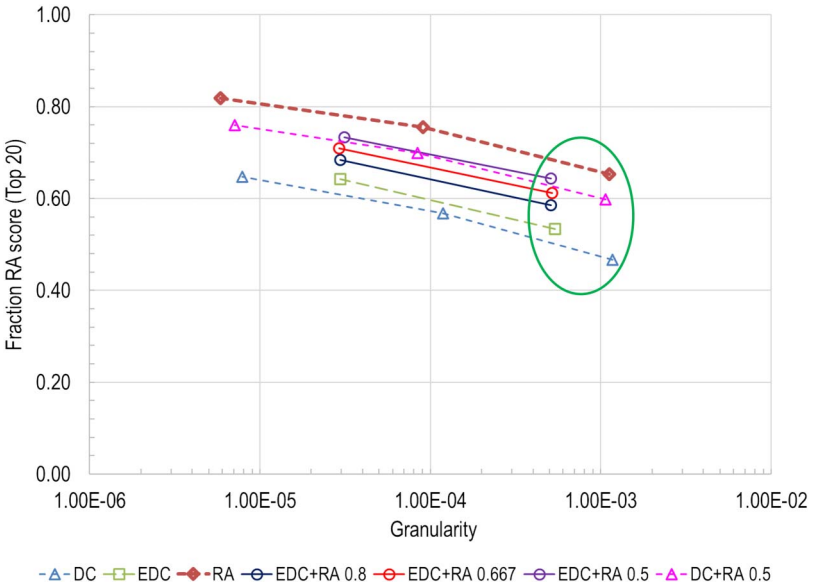

Figure 4. RA score fractions for the seven PubMed models.

this metric proportionally as citation information is added. EDC performs slightly better than DC using this metric, which suggests that the additional citation information in EDC contributes signal that improves performance.

Finally, Figure 5 shows that the performance differences between the different models are very modest using the grant link-based metric. The DC+RA model has the highest value, but not by much.

When all three metrics are considered together, it appears that the hybrid models outperform the DC, EDC, and RA models. Reliance on visual analysis is tenuous, however, as the granularities of the different solutions are different, making it difficult to determine which solution is best overall. Thus, we introduce here a single numerical value per model for each of the accuracy metrics to account for their different granularities.

For the Herfindahl index we applied curve fitting to the EDC curve in Figure 3 and use the resulting equation ($Herf_{EDC} = -0.0313*ln(gran) - 0.0843$, $R^2 = 0.998$) as a baseline. This equation approximates the Herfindahl index value for the EDC solution for different granularity values. For each of the other models (M), we estimated the ratio of its Herfindahl index value to that of the EDC model as

$$Rel\ Herf_M = Herf_M / Herf_{EDC}, \tag{10}$$

where $Herf_{EDC}$ is estimated using the same granularity value at which $Herf_M$ is measured. For example, $Herf_{DC}$ from Figure 3 is 0.1234 at a granularity of 0.00117161, while the estimated $Herf_{EDC}$ at this granularity is 0.1269. Using Eq. (10), the relative Herfindahl value for the DC model is 0.972. Values for each model are given in Table 2.

Similar calculations were done for the other two metrics. The RA curve was used as the baseline for the text-based metric (Figure 4) with a resulting curve fit equation $FracRA_{RA} = -0.0315*ln(gran) + 0.4467$, $R^2 = 0.974$. The EDC curve was used as the baseline for the grant-base metric (Figure 5) with a resulting curve fit equation of $FracGR_{EDC} = -0.0203*ln(gran) - 0.0975$, $R^2 = 1.00$, respectively. The relative values of the three metrics were averaged for each

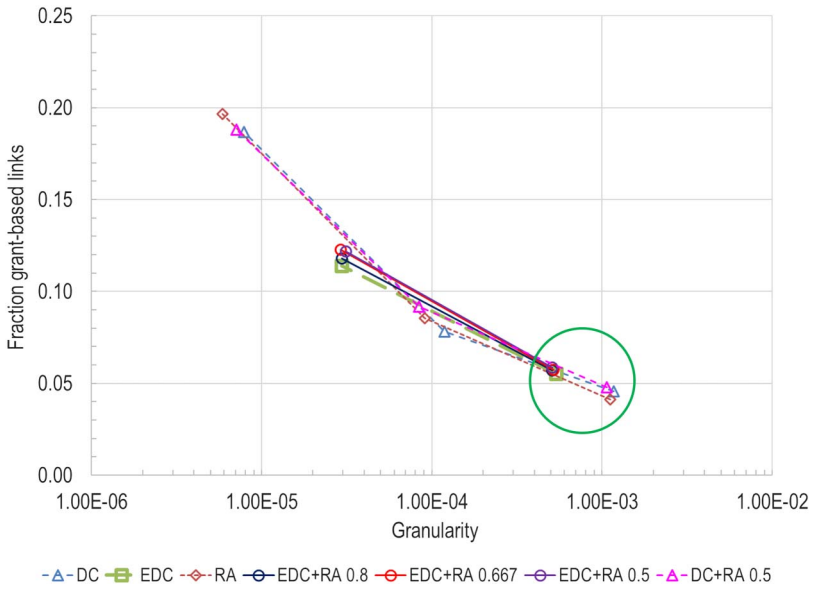

Figure 5. Fraction of grant-related links for the seven PubMed models.

Table 2. Relative accuracies of seven different models of PubMed documents

				EDC+RA			DC+RA
	DC	**EDC**	**RA**	**0.8**	**0.667**	**0.5**	**0.5**
Rel Herf	0.972	1.000	0.648	0.996	0.982	0.956	0.928
Rel RA	0.707	0.780	0.988	0.854	0.894	0.938	0.903
Rel GR	1.156	1.000	1.024	1.013	1.027	1.043	1.159
Average	0.945	0.927	0.887	0.954	0.968	0.979	**0.997**

model (see Table 2) and show that the hybrid models do indeed outperform the DC, EDC, and RA models. The DC+RA model is best overall and is significantly better than the RA and DC models to which it can be directly compared.

Surprisingly, DC outperformed EDC, which is in partial conflict with the results of other recent studies. The differences in results may be due to the different metrics that were used as bases of comparison. Ahlgren et al. (2020) showed that EDC is far more accurate than DC when comparing using a sophisticated MeSH-based metric. Waltman et al. (2020) found that EDC is more accurate than DC when compared using the BM25 text metric. Table 2 shows that our experiments agree with these results when EDC is compared with DC using the RA text metric. However, Table 2 also shows that DC is more accurate than EDC when compared using the grant link-based metric, and that the difference was large enough to place DC above EDC using the averaged metric.

5.2. Model Overlaps

We were also interested in exploring the magnitude of the differences between models. How different is a model based on text from a model based on direct citation? To quantify the differences, we calculated the Adjusted Rand Index (ARI) between pairs of models. A similar type of analysis was performed by Velden et al. (2017) using normalized mutual information rather than ARI. ARI is a standard calculation to measure the similarity between two cluster solutions of the same set of objects. The standard Rand Index (RI) is the number of agreements divided by the

Table 3. Adjusted Rand Index between pairs of models

				EDC+RA			DC+RA
	DC	**EDC**	**RA**	**0.8**	**0.667**	**0.5**	**0.5**
DC							
EDC	0.457						
RA	0.289	0.261					
EDC+RA 0.8	0.432	0.699	0.298				
EDC+RA 0.667	0.423	0.647	0.334	0.744			
EDC+RA 0.5	0.391	0.570	0.376	0.665	0.734		
DC+RA 0.5	0.555	0.438	0.469	0.473	0.504	0.514	

Table 4. ARI values based on random rearrangements

Reassigned	ARI	Reassigned	ARI
1.0%	0.980	20%	0.640
2.0%	0.960	30%	0.490
5.0%	0.902	40%	0.360
10.0%	0.810	50%	0.258

number of agreements and disagreements in two cluster solutions. Agreements are defined as object pairs that are in the same subset in both solutions and those that are in different subsets in both solutions. Disagreements are those pairs of objects that are together in one cluster solution and not in the other. The ARI corrects for chance in the RI. This is particularly important when dealing with small numbers of objects, but in practice the ARI and RI are nearly identical for very large data sets. ARI values for pairs of our seven models are shown in Table 3 and range from 0.261 (between RA and EDC) to 0.744 (between EDC+RA 0.8 and EDC+RA 0.667).

One problem with the ARI is that we don't know exactly what the value means in this context. Just exactly how good is an ARI of 0.744? To answer this question, we took each of the seven models, selected a random 1% subset of documents, and randomly shuffled their cluster assignments, thus maintaining the same cluster sizes. ARI was then calculated between the original and modified cluster solutions. This was done at many different percentage levels from 1% to 50%. Table 4 shows the ARI associated with random rearrangements of different magnitudes. Each value is the average of between 10 and 15 separate calculations. Standard deviations appear only in the fourth decimal place. The third-order polynomial fit to these values is $\%\text{Rearrange} = -0.2522\ \text{ARI}^3 + 0.7637\ \text{ARI}^2 - 1.2901\ \text{ARI} + 0.7782$ ($R^2 = 0.9999$), which allows us to estimate the level of rearrangement associated with any ARI value.

Estimated levels of rearrangements based on the ARI values are given in Table 5. These are estimates of the magnitude of the differences in paper-level assignments between pairs of models and are surprisingly large, ranging from a difference of 13.7% between EDC+RA 0.8 and EDC+RA 0.667 to 48.9% between RA and EDC. The differences of nearly 50% between the model based solely on text (RA) and the models based solely on citations (DC and EDC) are evidence of a fundamental gap between these two logics for measuring relatedness.

Table 5. Estimated level of rearrangements between pairs of models

				EDC+RA			DC+RA
	DC	EDC	RA	0.8	0.667	0.5	0.5
DC							
EDC	32.4%						
RA	46.3%	48.9%					
EDC+RA 0.8	34.3%	16.3%	45.5%				
EDC+RA 0.667	35.0%	19.5%	42.3%	13.7%			
EDC+RA 0.5	37.6%	24.4%	38.8%	18.4%	14.3%		
DC+RA 0.5	25.4%	33.8%	31.5%	31.3%	28.9%	28.3%	

6. DISCUSSION

In this study we compared PubMed models created using seven different relatedness measures: two based on direct citation, one based on text, and four using text+citation hybrid measures. We found that the hybrid relatedness measures outperform those based solely on text or direct citation and that the DC+RA model was the most accurate overall. In addition, the cluster size distribution for the DC+RA model is relatively flat, as opposed to the ECD+RA models, which have a steeper distribution. Although we have not developed methods to quantify the difference, our experience is that having clusters that are not too large and having a relatively large number of clusters with at least 200 documents is preferable to the alternatives for practical cluster-level analyses.

We also found that a hybrid model based on direct citation (DC) performs slightly better than one based on extended direct citation (EDC), and that different bases of comparison may give different overall results. Our previous models based solely on text or direct citation have proven to be very useful in practical contexts (Klavans & Boyack, 2017a). Although a hybrid model performs better numerically, it remains to be seen whether that will translate into increased usefulness in practice when used by decision makers.

Despite the relatively small differences in accuracy between models, there are large differences in the actual cluster contents. The fact that large differences in cluster contents are not accompanied by similarly large differences in relative accuracy suggests that there is no single best clustering (Gläser, Glänzel, & Scharnhorst, 2017), but that different logics prevail when using different relatedness measures, and that there is merit in each of those multiple logics. One challenge going forward will be to understand the relative strengths and weaknesses of each logic (text and citation) and use that knowledge to produce ever more accurate and useful practical models of science.

As a first step toward that end, additional analysis was done to identify which logic—textual or citation—was dominant in each cluster in the DC+RA hybrid model. To do this, we calculated

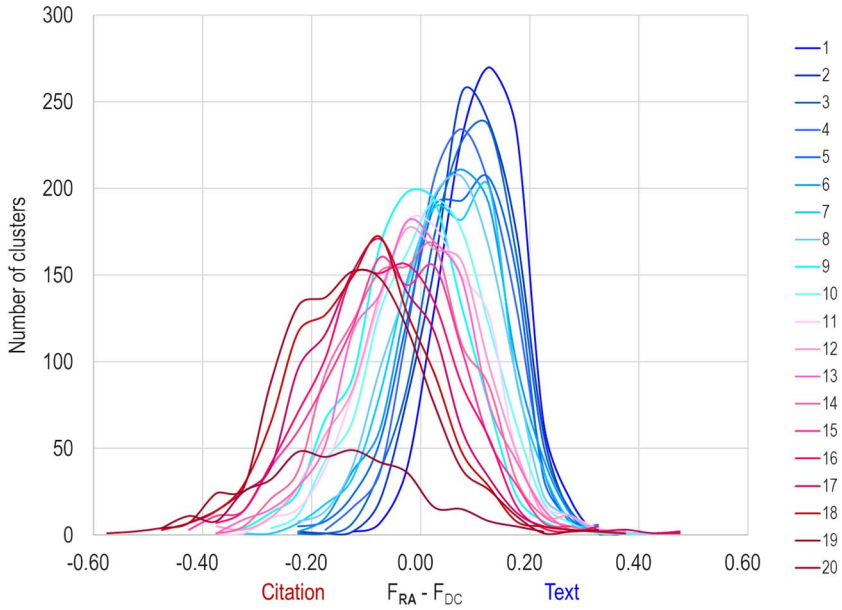

Figure 6. Distribution of clusters in the DC+RA model as a function of textual or citation dominance. Bins 1–19 contain 1,000 clusters each, while bin 20 contains 348 clusters.

the fraction of the original citation (DC) signal that was preserved within each cluster k and subtracted that from the fraction of the original textual (RA) signal that was likewise preserved.

$$F_{RA-k} - F_{DC-k} = \left(\sum r^{RA}_{ij-k} a_{ij} / \sum r^{RA}_{ij-k}\right) - \left(\sum r^{DC}_{ij-k} a_{ij} / \sum r^{DC}_{ij-k}\right), \quad (11)$$

where $a_{ij} = 1$ if documents i and j are in the same cluster and 0 if not. The resulting differences are plotted using bins of 1,000 clusters where clusters are ordered by descending size. Bin 1 contains the 1,000 largest clusters, and so on. Figure 6 shows that a majority of the largest clusters (blue lines) preserve more textual signal than citation signal, while a majority of the smaller clusters (red lines) preserve more citation signal than textual signal. Thus, textual logic seems to play a larger role in the identification of larger clusters, while citation logic seems to play a larger role in the identification of smaller clusters. This is very interesting behavior that will require further investigation to untangle.

Finally, it is useful to once again point out that the large-scale models in this study were based on open data. PubMed RA scores have been found to yield relatively accurate results (Boyack et al., 2011), are precomputed, and, although limits on numbers of queries per minute apply, can be freely downloaded. The new NIH OCC is large and robust, is currently being updated frequently, and while it does not cover every PubMed document, is sufficiently broad to be used for bibliometrics purposes. Given the open availability of these data, we recommend that those doing small-scale bibliometric studies in biomedical fields should upgrade their efforts and make use of the large-scale resources that are now freely available.

ACKNOWLEDGMENTS

We appreciate comments from Caleb Smith on working drafts of this paper.

AUTHOR CONTRIBUTIONS

Kevin Boyack: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Richard Klavans: Conceptualization, Writing—original draft.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This work was supported by NIH award HHSN271201800033C.

DATA AVAILABILITY

All data used in this study are openly available, either from PubMed or through one of the data sources or utilities referenced in the footnotes.

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. **DOI:** https://doi.org/10.1162/qss_a_00027

Ahlgren, P., & Colliander, C. (2009). Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, *3*, 49–63. **DOI:** https://doi.org/10.1016/j.joi.2008.11.003

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404. **DOI:** https://doi.org/10.1002/asi.21419

Boyack, K. W., & Klavans, R. (2018). Accurately identifying topics using text: Mapping PubMed. *23rd International Conference on Science and Technology Indicators (STI 2018)*. Retrieved from http://hdl.handle.net/1887/65319

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., … Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLOS ONE*, *6*(3), e18029. **DOI:** https://doi.org/10.1371/journal.pone.0018029, **PMID:** 21437291, **PMCID:** PMC3060097

Glänzel, W., & Thijs, B. (2011). Using 'core documents' for the representation of clusters and topics. *Scientometrics*, *88*, 297–309. **DOI:** https://doi.org/10.1007/s11192-011-0347-4

Glänzel, W., & Thijs, B. (2017). Using hybrid methods and 'core documents' for the representation of clusters and topics: The astronomy dataset. *Scientometrics*, *111*, 1071–1087. **DOI:** https://doi.org/10.1007/s11192-017-2301-6

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data– different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, *111*, 981–998. **DOI:** https://doi.org/10.1007/s11192-017-2296-z

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*, 436–447. **DOI:** https://doi.org/10.1016/j.joi.2018.03.004

Hutchins, B. I., Baker, K. L., Davis, M. T., Diwersy, M. A., Haque, E., … Santangelo, G. M. (2019). The NIH Open Citation Collection: A public access, broad coverage resource. *PLOS Biology*, *17*(10), e03000385. **DOI:** https://doi.org/10.1371/journal.pbio.3000385, **PMID:** 31600197, **PMCID:** PMC6786512

Janssens, F., Quoc, V. T., Glänzel, W., & de Moor, B. (2006). Integration of textual content and link information for accurate clustering of science fields. *International Conference on Multidisciplinary Information Sciences and Technologies (InSciT2006)* (pp. 615–619).

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, *45*, 683–702. **DOI:** https://doi.org/10.1016/j.ipm.2009.06.003

Klavans, R., & Boyack, K. W. (2017a). Research portfolio analysis and topic prominence. *Journal of Informetrics*, *11*(4), 1158–1174. **DOI:** https://doi.org/10.1016/j.joi.2017.10.002

Klavans, R., & Boyack, K. W. (2017b). 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

Lin, J., & Wilbur, W. J. (2007). PubMed related articles: A probabilistic topic-based model for content similarity. *BMC Bioinformatics*, *8*, 423. **DOI:** https://doi.org/10.1186/1471-2105-8-423, **PMID:** 17971238, **PMCID:** PMC2212667

Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., & De Moor, B. (2010). Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the American Society for Information Science and Technology*, *61*(6), 1105–1119. **DOI:** https://doi.org/10.1002/asi.21312

Meyer-Brötz, F., Schiebel, E., & Brecht, L. (2017). Experimental evaluation of parameter settings in calculation of hybrid similarities: Effects of first- and second-order similarity, edge cutting, and weighting factors. *Scientometrics*, *111*, 1307–1325. **DOI:** https://doi.org/10.1007/s11192-017-2366-2

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

Sjögårde, P., & Ahlgren, P. (2020). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. *Quantitative Science Studies*, *1*(1), 207–238. **DOI:** https://doi.org/10.1162/qss_a_00004

Thijs, B., Schiebel, E., & Glänzel, W. (2013). Do second-order similarities provide added-value in a hybrid approach? *Scientometrics*, *96*, 667–677. **DOI:** https://doi.org/10.1007/s11192-012-0896-1

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*, 5233. **DOI:** https://doi.org/10.1038/s41598-019-41695-z, **PMID:** 30914743, **PMCID:** PMC6435756

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

Waltman, L., Boyack, K. W., Colavizza, G., & Van Eck, N. J. (2017). A principled methodology for comparing relatedness measures for clustering publications. Paper presented at the 16th International Conference of the International Society on Scientometrics and Informetrics, Wuhan, China.

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. **DOI:** 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. **DOI:** https://doi.org/10.1002/asi.22748

Yu, D., Wang, W., Zhang, S., Zhang, W., & Liu, R. (2017). Hybrid self-optimized clustering model based on citation links and textual features to detect research topics. *PLOS ONE*, *12*(10), e0187164. **DOI:** https://doi.org/10.1371/journal.pone.0187164, **PMID:** 29077747, **PMCID:** PMC5659815

Zhang, L., Glänzel, W., & Ye, F. Y. (2016). The dynamic evolution of core documents: An experimental study based on h-related literature (2005–2013). *Scientometrics*, *106*, 369–381. **DOI:** https://doi.org/10.1007/s11192-015-1705-4