Accurately identifying topics using text: Mapping PubMed

Boyack, K.W.; Klavans, R.

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Accurately identifying topics using text: Mapping PubMed[1]

Kevin W. Boyack* and Richard Klavans**

* *kboyack@mapofscience.com*
SciTech Strategies, Inc., Albuquerque, NM 87122 (USA)

** *rklavans@mapofscience.com*
SciTech Strategies, Inc., Wayne, PA 19087 (USA)

Introduction

Recently, citation links have been shown to produce accurate delineations of tens of millions of scientific documents into a large number (~100,000) of clusters (Sjögårde & Ahlgren, 2018). Such clusters, which we refer to as topics, can be used for research evaluation and planning (Klavans & Boyack, 2017a) as well as to identify hot and/or emerging topics (Small, Boyack, & Klavans, 2014). While direct citation links have been shown to produce more accurate topics using large citation databases than co-citation or bibliographic coupling links (Klavans & Boyack, 2017b), no such comparison has been done at a similar scale using topics based on textual relatedness due to the extreme computational requirements of calculating an enormous number of document-document similarities using text. Thus, we simply do not know if topics identified from a large database using textual characteristics are as accurate as those that are identified using direct citation. This paper aims to fill that gap. In this work we cluster over 23 million documents from the PubMed database (1975-2017) using a text-based similarity and compare the accuracy of the resulting topics to those from existing citation-based topics using three different measures.

Background

Two lines of recent research inform the current work. First, clustering routines are now available that can cluster tens of millions of documents. Waltman & van Eck made a key advance by introducing a modularity-based clustering algorithm whose utility was demonstrated by clustering 10 million WOS documents using 98 million direct citation links (Waltman & van Eck, 2012). They also introduced a second modularity-based algorithm, known as the smart local moving (SLM) approach (Waltman & van Eck, 2013). This algorithm has quickly become the algorithm of choice for those clustering large document sets (Klavans & Boyack, 2017b; Ruiz-Castillo & Waltman, 2015), and has been used to cluster up to 60 million documents with 800 million edges by the current authors (unpublished). Note that while the SLM approach has primarily been used with direct citations, it can be used to cluster documents (or other objects) using any type of relatedness value. Direct citations are not required.

Second, there has been an increased focus on the accuracy of cluster solutions. The first large scale comparison of the accuracy of clusters resulting from different relatedness measures was

[1] This work was supported by NIH award HHSN271201700041C.

that done by Boyack and colleagues (Boyack & Klavans, 2010; Boyack et al., 2011). Using a set of 2.15 million PubMed/Scopus documents where document-document relatedness was computed 13 different ways (3 citation-based, 9 text-based, 1 hybrid), they found that the best citation-based and best text-based approaches gave results of similar accuracy. Among text-based approaches, the PubMed related articles (RA) measure gave the best results when accuracy was calculated using the concentration of grant-article linkages. More recently, and using over 40 million documents from Scopus, Klavans & Boyack (2017b) compared accuracies as a function of cluster size for clusters created using direct citation, co-citation, and bibliographic coupling using the SLM approach, finding that direct citation gave the best results. This work was followed up by Sjögårde & Ahlgren (2018) who, using the SLM approach and direct citation on 31 million WOS documents, identified a granularity which maximized topic accuracy as measured using an adjusted Rand index.

Data and Methods

For this work we chose to use the PubMed database for two reasons: 1) PubMed is a large database that is known to have both broad and deep coverage of the biomedical literature, and 2) RA scores for pairs of documents have been pre-calculated by the U.S. National Library of Medicine (NLM), thus alleviating the need to calculate document-document relatedness values. RA scores are based on the words (terms) in titles and abstracts, along with MeSH terms, where scores are calculated as the sum of the weights (local wt1 × local wt2 × global wt) over all terms using local weights computed using the algorithm of Lin & Wilbur (2007).[2] NLM processes each new record added to the database using the related articles algorithm to identify up to 100 similar articles along with their scores. With each calculation, reciprocal links are also added between the older records and new record. Thus, older records may have significantly more than 100 similar articles. We assume that the 100 RA scores computed for each new record are the "top 100" such scores. However, we have not been able to verify this with NLM staff.

Our PubMed database was updated on August 31, 2017 to include all records from 1960 forward. Our store of RA scores was also updated shortly thereafter,[3] and contains scores for 4.38 billion document pairs in total. Figure 1 shows the number of PubMed records per year along with the percentage of those records containing various metadata, such as abstracts, MeSH terms, and funding acknowledgments. Although MeSH terms have been assigned for a large majority of documents, few PubMed records contained abstracts prior to 1975. Given that abstracts are a large contributor to the RA scores, we chose to limit our clustering to the 23,234,923 PubMed records from 1975 forward for which we had RA scores.

Clustering of PubMed documents and subsequent characterization of accuracy was accomplished using the following process:

1) Two different filtered subsets of the RA scores were created, one using the top 12 scores for each document (12PM) and the other using the top 40 scores for each document (40PM).
2) Each of the two resulting similarity files were then clustered using the SLM algorithm. We desired cluster solutions at three different levels of granularity – approximately 50000, 5000, and 500 clusters – to give cluster size distributions that would be consistent with those in our existing direct citation solutions (Klavans & Boyack, 2017b) to enable reasonable comparisons. Resolution values were chosen to achieve these results, as shown in Table 1. SLM runs were done using 10 iterations.

[2] https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Computation_of_Similar_Articl
[3] Scores can be retrieved using Entrez queries, see https://www.ncbi.nlm.nih.gov/books/NBK25499/.

3) Results for each of our two models (12PM and 40PM) were then compared with those from the existing citation-based solutions using three different measures of relative accuracy (that will be explained later).

Figure 1: Numbers of PubMed records per year along with the percentage that contain abstracts, MeSH terms, general funding acknowledgments, and specific grant information.

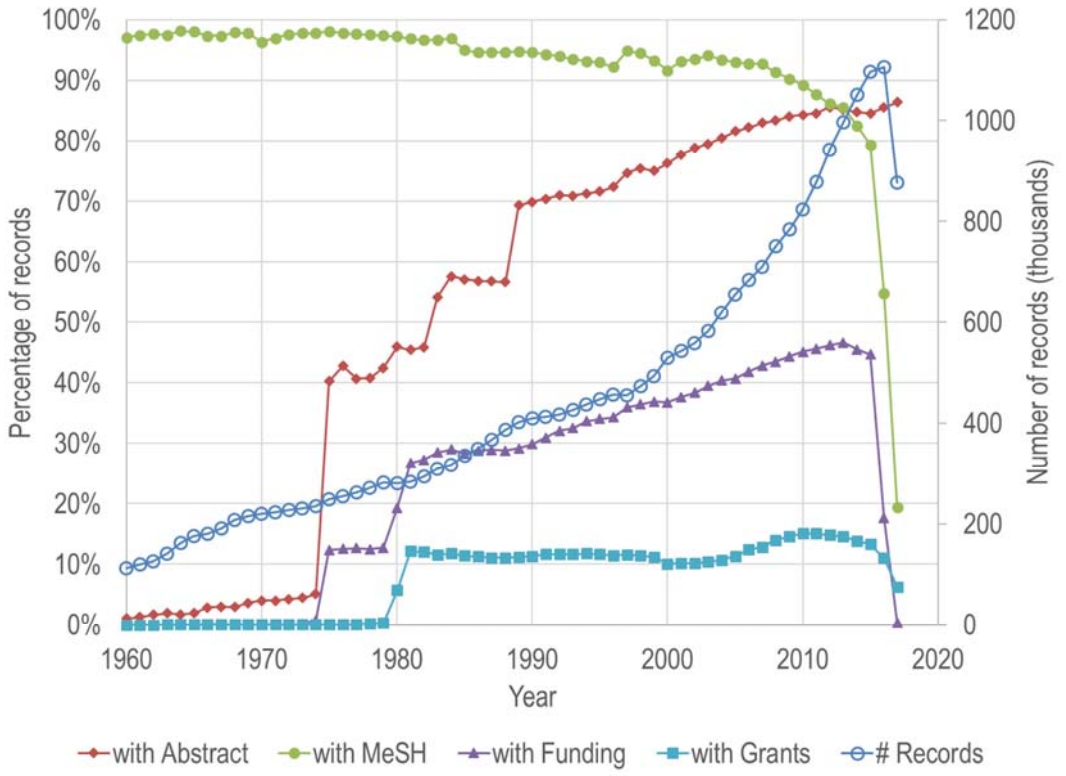


Table 1. Input resolution values (italicized) for the SLM clustering runs.

Input subset	**~50,000 clusters**	**~5,000 clusters**	**~500 clusters**
12PM (241,643,546 pairs)	*12787.5*	*4.25×10^{8}*	*2.0×10^{9}*
40PM (790,047,772 pairs)	*23250*	*2.5625×10^{9}*	*2.5625×10^{10}*
Min cluster size	30	300	3000

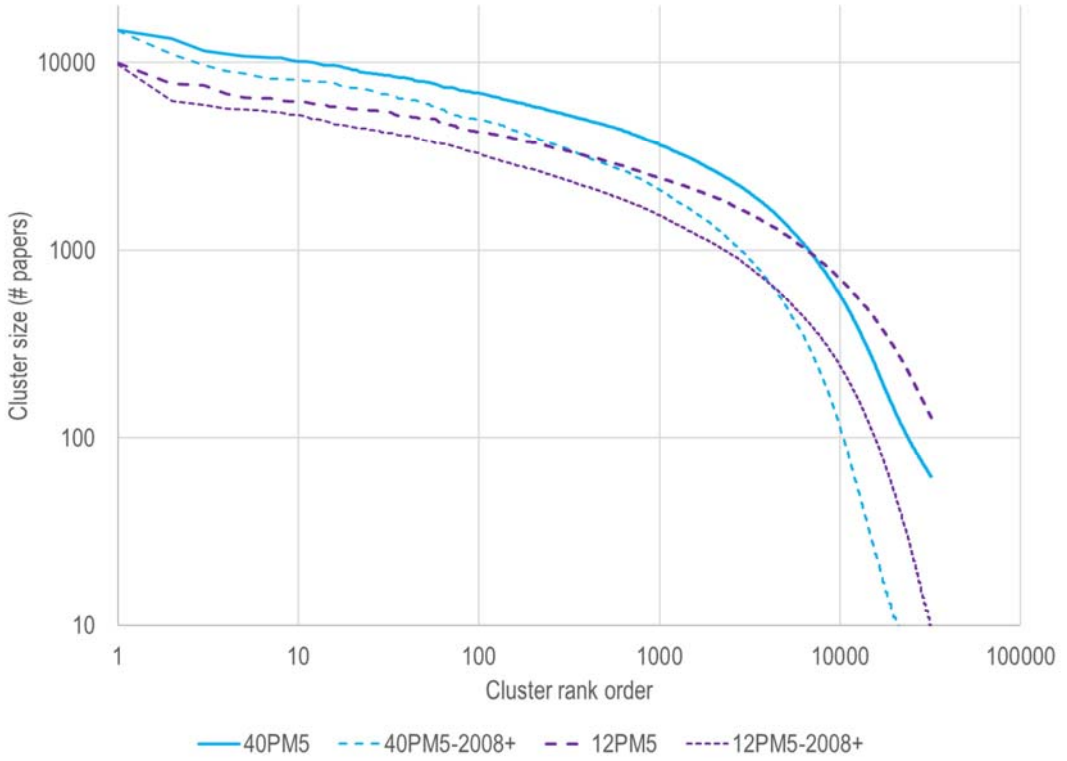
Results

Table 2 shows characteristics of both models at the 50,000 cluster level, along with numbers of clusters above several size thresholds. The $12PM^5$ and $40PM^5$ models are quite different. Although they have a similar number of clusters overall, the $40PM^5$ model has larger clusters and significantly fewer clusters with at least 100 and 500 papers, as also shown by the cluster size distributions in Figure 2. Although both models have around the same number of clusters with at least 1000 papers, the $12PM^5$ model has far more clusters with between 100 and 1000 papers. We favor models with larger numbers of such clusters in that they correspond well with the perceptions of researchers.

Table 2. Characteristics of models with ~50,000 clusters (superscript[5] denotes ~10^5 clusters).

Model	# Papers	# Clust	# Clust >30	# Clust >100	# Clust >500
12PM[5]	23,234,841	55,202	54,018	36,466	13,829
40PM[5]	23,234,923	53,819	46,067	23,855	10,897

Figure 2: Size distributions for the 12PM[5] and 40PM[5] clusters for all years and since 2008.



In accordance with best practices, we seek to establish the relative accuracy of the models created in this project by comparing them with existing models of known accuracy. Such comparisons should be done using a principled approach (Waltman, Boyack, Colavizza, & Van Eck, 2017). For example, solutions should be compared using a basis that is independent of all approaches if possible. In practice, this condition is difficult to achieve. Furthermore, solutions with different granularities (i.e., different numbers and sizes of clusters) should not be directly compared because most metrics will naturally favor solutions with few large clusters (low granularity) over those with many small clusters (high granularity). Granularity should be accounted for in any comparative analysis.

In practice, it is nearly impossible to generate multiple solutions with the same granularity. We overcome this by using the graphical approach of Waltman et al. (2017) that presents results as granularity-accuracy (GA) plots, where granularity is defined as the number of papers in the solution divided by the sum of the squared cluster sizes.

Here we compare the accuracy of our PubMed-based models to several citation-based models from our earlier research (Klavans & Boyack, 2017b), including direct citation, bibliographic coupling, and co-citation models created using Scopus data. These comparisons are based on

only those documents that are available in both Scopus and PubMed. We restrict the comparison in this way to avoid biasing the results by including content not available in PubMed. Three separate measures of relative accuracy are used: 1) a citation-based measure, which will likely bias toward citation-based models, 2) a text-based measure, which will likely bias toward text-based models, and 3) a grant linkage-based measure, which is relatively independent of both citation-based and text-based models. While each of these measures provides a discrete perspective regarding accuracy, we acknowledge that there is no absolute ground truth. That is why we use multiple measures, to reflect the fact that different organizing logics exist in science and that different relatedness measures and clustering techniques may be more or less aligned with different logics.

Citation-based measure

Our first comparison measure is a concentration index using the references from review articles. The premise behind this comparison is that those writing review articles are proficient in their topics, and that each review thus serves as an expert opinion of the contents of a topic in science. Under this assumption, the model that most closely duplicates the topic structure suggested by thousands of reviews is the most accurate.

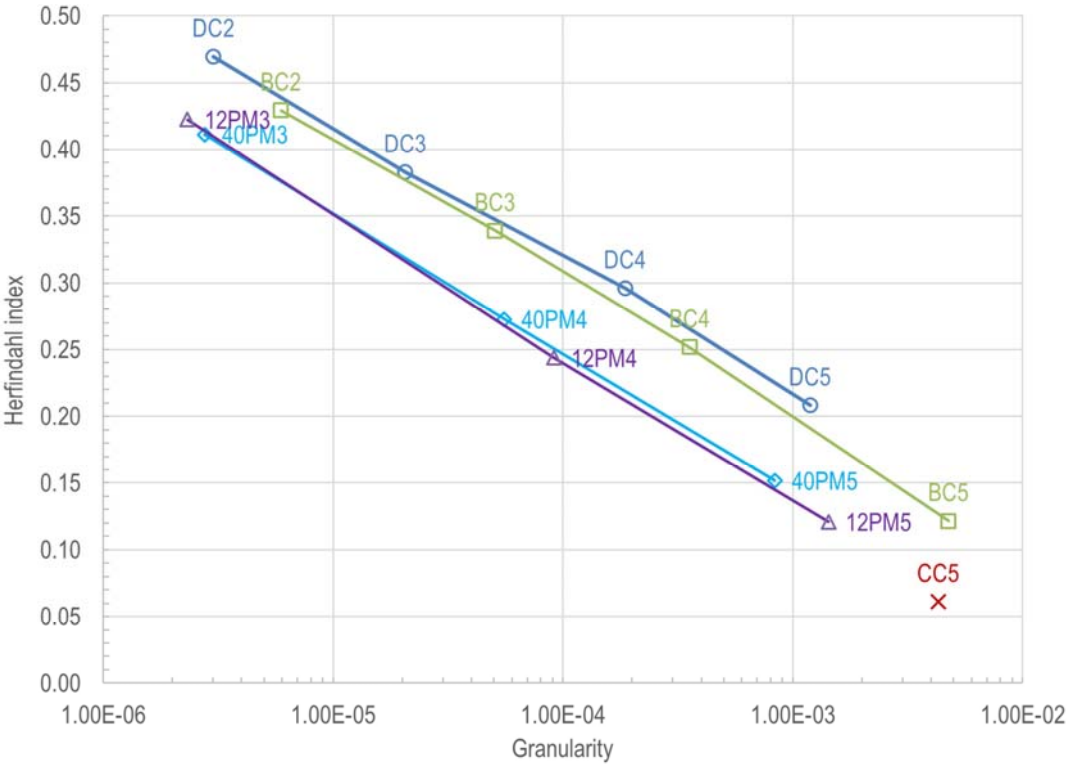
We identified 17,873 papers from PubMed published in 2010 that had at least 100 references each, and for which at least 80% of the references were available in PubMed and Scopus. The reference information was obtained by matching PubMed and Scopus records, both for the review papers and for the references. Herfindahl index values were calculated for each review paper and model using a method described previously (Klavans & Boyack, 2017b). The overall Herfindahl index value for a model is the average of the index values for the 17,873 individual papers. Figure 3 shows a comparison of these values for our PubMed models and several citation-based models.

The figure shows results for models with a wide range of granularities for larger context, where the slope of the lines is important in that it shows the natural relationship between granularity and reference concentration. We are most interested in the results with a granularity of around 10^{-3}, which is the level most applicable to researchers. Here, direct citation (DC^5) and bibliographic coupling (BC^5) models have higher index values than the text-based models ($12PM^5$, $40PM^5$), as expected, and thus are more accurate from the point of view of this measure. When taking slopes into consideration, the $12PM^5$ and $40PM^5$ models have very similar accuracy. Also, the difference between DC^5 and $12PM^5$ models is perhaps not as large as the figure might suggest. The Herfindahl index is based on squared values which accentuates differences between curves. Thus, differences in accuracy between models based on Figure 3 should not be judged in a linear sense.

Text-based measure

Our second comparison measure is the textual RA scores themselves. This measure is not independent of the models, having been used to cluster the 12PM and 40PM models, and does not provide a fair comparison between text-based and citation-based models. Nevertheless, we include it for two reasons. First, the Herfindahl index is inherently biased to citation-based similarities. Thus, we include this text-biased measure for balance. Second, since this measure was used for clustering as well as for analysis, it provides an upper bound to accuracy based on textual characteristics (Waltman et al., 2017).

Figure 3: Herfindahl index as a function of granularity. DC = direct citation, BC = bibliographic coupling, CC = co-citation, 12PM, 40PM = PubMed RA top 12 and top 40.



To calculate this metric, we take the 20 highest scoring RA pairs for each article from 1996 to 2012. Pairs where the linked document was outside this time window were later excluded. Thus, for many papers, fewer than 20 similarity pairs were included in the basis set, which ultimately consisted of 203,537,938 pairs of documents with PMRA scores. For each model, the metric value was then calculated as the sum of the scores for paper pairs that were in the same cluster divided by the total sum of the scores. This metric thus reflects the fraction of the overall textual signal that ends up within clusters.

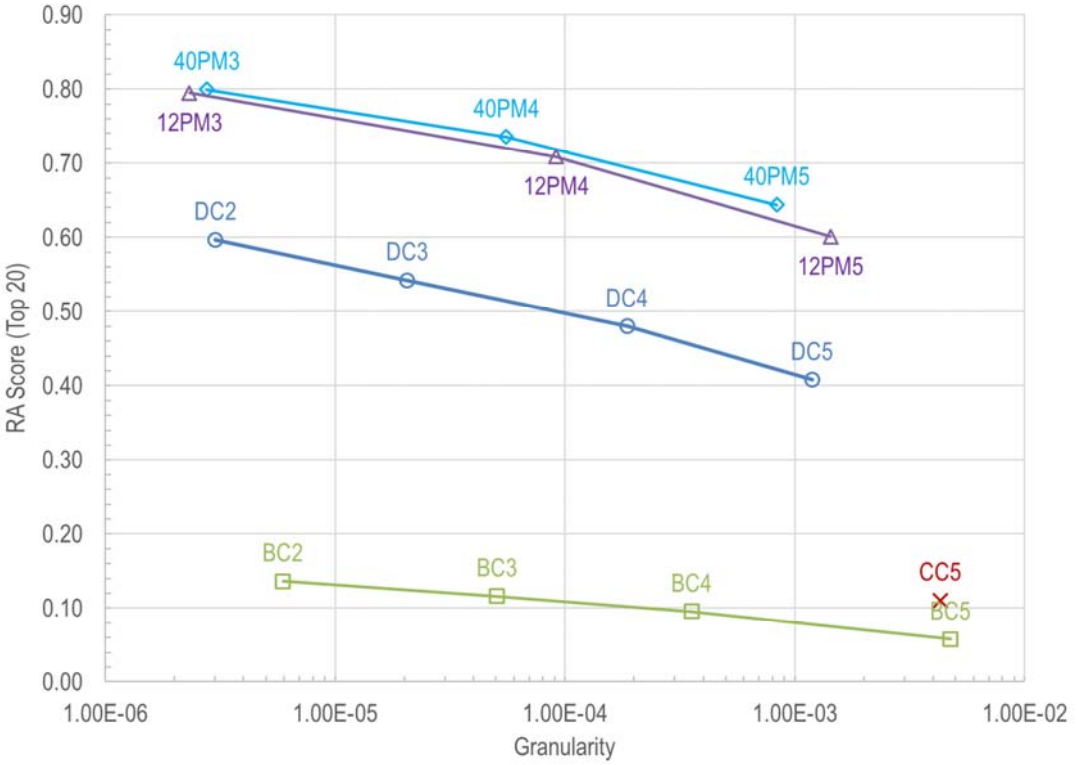
Results for this metric for all models are shown in Figure 4. As expected, the two sets of models based on RA scores do the best job of creating clusters where most of the text similarity is within clusters. The $40PM^5$ and $12PM^5$ models both preserve over 60% of the overall textual similarity within clusters, while the DC^5 model only preserves 40% of the overall textual similarity within clusters. We note that, while the $40PM^5$ has the best overall score among models with granularity of around 10^{-3}, it is not substantially higher than the score for the $12PM^5$ model. Thus, including the additional similarities (from top 12 to top 40) does not appear to substantially increase the accuracy of the model.

Grant linkage-based measure

Our third comparison measure is based on grant-article linkages mined from the acknowledgments of papers. From these we create a list of pairs of papers that acknowledge the same grant, and then compute the fraction of those pairs of papers that end up in the same cluster for each model. This is different from, but related to, a measure we used in previous work (Boyack & Klavans, 2010; Boyack et al., 2011) where grant-article linkages were used to calculate a Herfindahl index based on cluster assignments. The logic behind this measure is

the assumption that two papers produced from the same grant should belong to the same topic. This grant-based measure is the most objective of our three measures in that it is completely independent from the data used to create the models under comparison.

Figure 4: Fraction of textual similarity preserved within clusters as a function of granularity.



Grant-article linkage data were obtained from the NIH RePORTER website comprising 3,686,992 links between PubMed article IDs and NIH project numbers for papers published from 1980 to 2012. From these links, we identified 263,334,240 pairs of papers that reference the same NIH project number. The fraction of these pairs of papers that appear in the same cluster has been calculated for each model and are shown in Figure 5. Among models with a granularity of around 10^{-3}, the DC^5, $12PM^5$ and $40PM^5$ models all do an equally good job of creating clusters that preserve grant-based relationships. From the grant-based perspective the three models thus have the same relative accuracy.

Discussion

With three different accuracy metrics, one unbiased, one biased toward citation-based similarities, and one biased toward text-based similarities, we now have enough information to draw observations about the relative accuracies of different models. Table 3 contains results comparing the DC^5, $12PM^5$ and $40PM^5$ models, and shows that overall, one can consider all three models to have comparable accuracy. Thus, we find that topics identified from a large database using a sophisticated textual relatedness measure are just as accurate as those that are identified using direct citation. This does not suggest that all text-based relatedness measures will do as well as the PubMed RA measure. For example, given our previous results (Boyack et al., 2011) we suspect that relatedness based on MeSH term co-occurrence will not suffice. Nevertheless, these results suggest that topics based on text can be

competitive with those based on citations at large scale, which is particularly important for work with open source databases and those that do not include cited references.

Figure 5: Fraction of paper pairs referencing the same NIH project number appearing within clusters as a function of granularity.

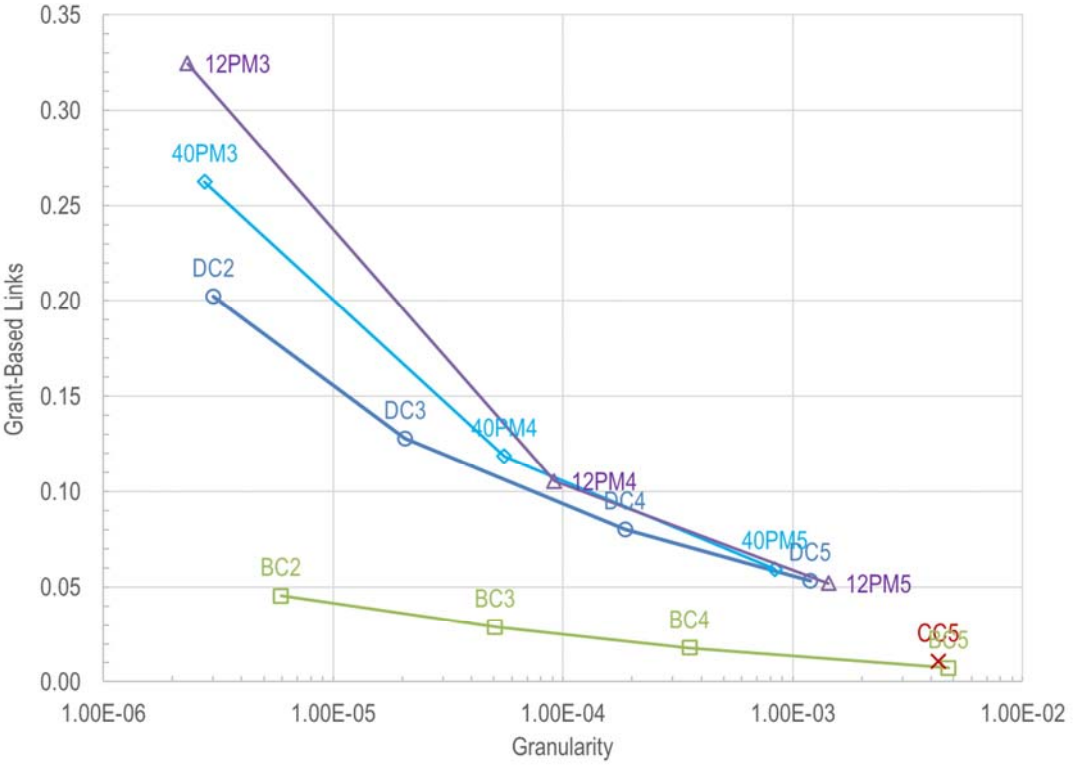


Table 3. Comparison of accuracy metrics for X^5 models.

Measure	**Result**
Citation-based measure	$DC^5 > (12PM^5, 40PM^5)$
Text-based measure	$(12PM^5, 40PM^5) > DC^5$
Grant-based measure	$(12PM^5, 40PM^5) = DC^5$
Overall	$(12PM^5, 40PM^5) \approx DC^5$

References

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.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., et al. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS One, 6*(3), e18029.

Klavans, R., & Boyack, K. W. (2017a). Research portfolio analysis and topic prominence. *Journal of Informetrics, 11*(4), 1158-1174.

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.

Lin, J., & Wilbur, W. J. (2007). PubMed related articles: A probabilistic topic-based model for content similarity. *BMC Bioinformatics, 8*, 423.

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics, 9*, 102-117.

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.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*, 1450-1467.

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

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.

Waltman, L., & van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B, 86*, 471.