

## Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?

Kevin W. Boyack<sup>a</sup> & Richard Klavans<sup>b</sup>

<sup>a</sup> SciTech Strategies, Inc., Albuquerque, NM 87122 USA ([kboyack@mapofscience.com](mailto:kboyack@mapofscience.com))

<sup>b</sup> SciTech Strategies, Inc., Berwyn, PA 19312 USA ([rklavans@mapofscience.com](mailto:rklavans@mapofscience.com))

### Abstract

In the past several years studies have started to appear comparing the accuracies of various science mapping approaches. These studies primarily compare the cluster solutions resulting from different similarity approaches, and give varying results. In this study, we compare the accuracies of cluster solutions of a large corpus of 2,153,769 recent articles from the biomedical literature (2004-2008) using four similarity approaches: co-citation analysis, bibliographic coupling, direct citation, and a bibliographic coupling-based citation-text hybrid approach. Each of the four approaches can be considered as a way to represent the research front in biomedicine, and each is able to successfully cluster over 92% of the corpus. Accuracies are compared using two metrics – within-cluster textual coherence as defined by the Jensen-Shannon divergence, and a new concentration factor based on the grant-to-article linkages indexed in MEDLINE. Of the three pure citation-based approaches, bibliographic coupling slightly outperforms co-citation analysis using both accuracy measures; direct citation is the least accurate mapping approach by far. The hybrid approach improves upon the bibliographic coupling results in all respects. We consider the results of this study to be robust given the very large size of the corpus, and the specificity of the accuracy measures used.

### Introduction

Science mapping has reached the point where it is no longer a primarily academic venture but instead is being driven by and used for practical purposes. Although such mapping is often equated with visual representations of the structure of science, the visuals are only a reflection of the layout and partitioning of bibliographic units (e.g., documents, words, authors, journals) that are the primary output of the mathematics behind the mapping. The partitions themselves, along with detailed analysis of the partitions, are typically of far more interest to decision-makers than are visuals of the structure. The accuracy of these partitions becomes very important, especially when these maps are used for real-world problems of research planning and evaluation.

Our work over the past several years has been aimed specifically at creating ever more detailed (Klavans & Boyack, 2010), accurate (Boyack, Klavans, & Börner, 2005; Klavans & Boyack, 2006a, 2006b), and actionable science maps that can be used by decision-makers. We have been very encouraged to see studies showing up in the recent literature aimed at increasing the accuracy of maps. Studies have compared citation approaches (Jarneving, 2005; Shibata, Kajikawa, Takeda, & Matsushima, 2009), citation with textual similarity approaches (Ahlgren & Jarneving, 2008; Glenisson, Glänzel, Janssens, & de Moor, 2005; Glenisson, Glänzel, & Persson,

2005), and several have even started to examine hybrid text-citation approaches (Ahlgren & Colliander, 2009a; Janssens, Quoc, Glänzel, & de Moor, 2006; Janssens, Zhang, De Moor, & Glänzel, 2009; Liu et al., 2010).

We recently completed a study for the U.S. National Institutes of Health (NIH) in which we compared science maps generated from a single large corpus (2.15 million documents published from 2004-2008) using 13 different similarity approaches, including three citation-based approaches, nine text-based approaches, and one hybrid approach. The ultimate application of this science mapping effort for NIH will be for portfolio planning and analysis; any time science mapping has the potential to become co-mingled with funding and decision making, the map must be as accurate as possible. Thus, our study was focused on determining how to generate the most accurate large scale map of the medical science literature for portfolio analysis applications.

In any study of accuracy, the question of how to measure and compare the accuracies of different solutions must be answered. When it comes to the accuracy of science maps this is not a trivial issue given that ground truth rarely exists. In this article we use two measures of accuracy to compare the cluster solutions from the different similarity approaches. The first is a textual coherence measure, and the second is a new measure that we introduce here. This new measure uses grant-to-article linkages from the acknowledgements of grants indexed in MEDLINE. Cluster solutions that provide a higher concentration of articles from single grants are considered more accurate than those with a lower concentrating factor. This measure is particularly well aligned with portfolio analysis since the portfolios that will be analyzed are most often tied to funded grants.

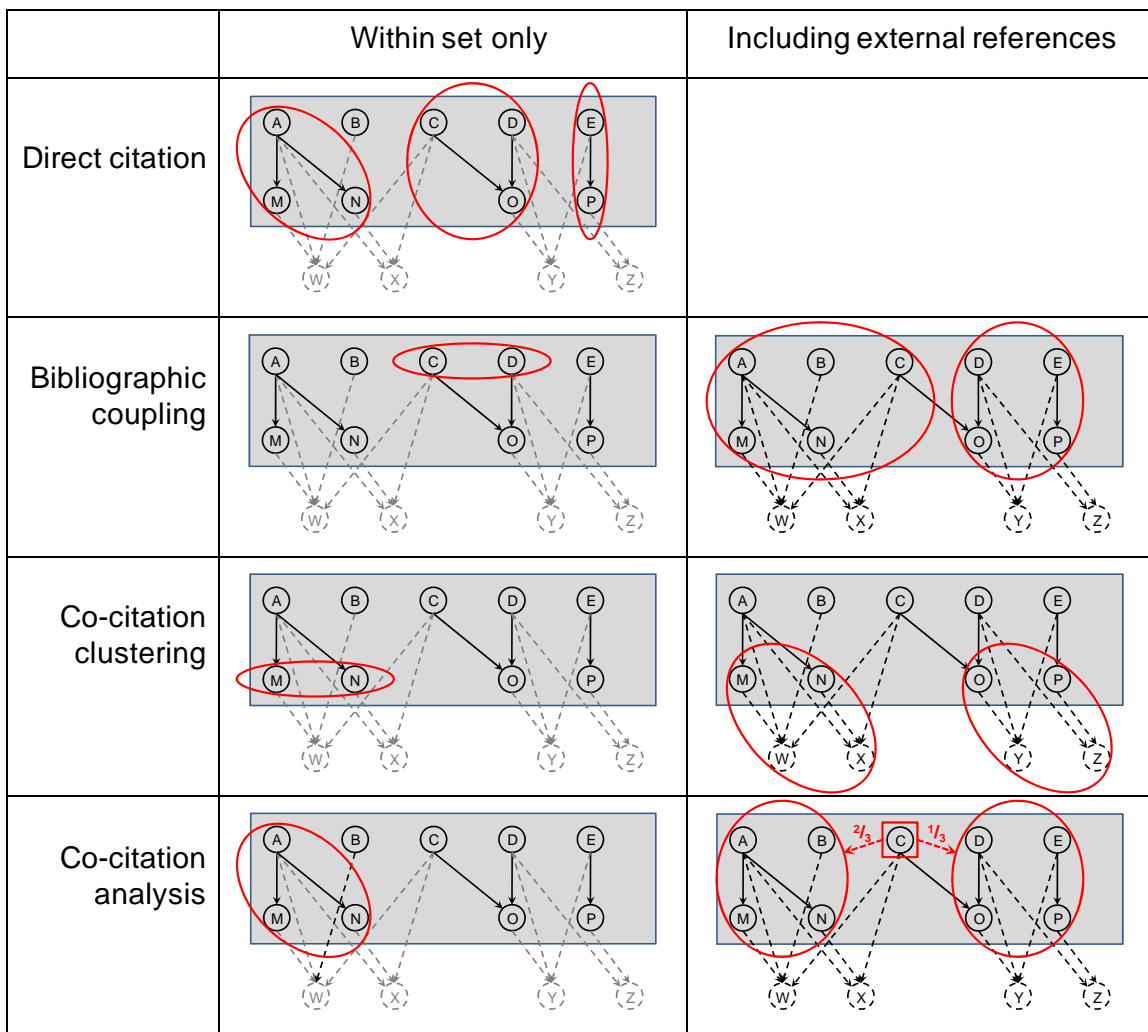
The entire NIH study was far too broad and varied to be fully reported on in a single article. Thus, this article is restricted to comparing science maps from three citation-based approaches – co-citation analysis, bibliographic coupling, and direct citation – and the hybrid approach. We include the hybrid approach here because it was based on bibliographic coupling. We acknowledge that text-based mapping methodologies are very well developed, and will report the results of the text-based approaches from our large study in other articles.

## **Background**

The three main citation-based approaches to science mapping have somewhat different histories. While bibliographic coupling (Kessler, 1963) and co-citation analysis (Marshakova, 1973; Small, 1973) both trace their roots back several decades, co-citation analysis was adopted as the *de facto* standard in the 1970s, and has enjoyed that position of preference ever since. There has been a recent resurgence in the use of bibliographic coupling that is challenging the historical preference for co-citation analysis (Boyack, 2009b; Jarneving, 2005, 2007a, 2007b; Sandström, 2009). Direct citation (sometimes also called intercitation), although employed from time to time (cf., Shibata, Kajikawa, Takeda, & Matsushima, 2008), has not been as widely used because of the need to use very long time windows to obtain a sufficient linking signal for clustering. Although these three approaches are typically not combined, Small (1997) did propose a way to effectively combine them, although no-one seems to have pursued use of this combined linkage technique at large scale.

### Citation-based mapping approaches

Before reviewing the history of accuracy studies associated with science mapping, we feel the need to clarify the differences between citation-based mapping approaches, and how they each cover the document space and represent the research front. Figure 1 shows a longitudinal data set containing nine articles (gray box) – articles A-E are recently published and not yet cited, while articles M-P are older, and have been cited by the more recent documents in the set. In addition, we show articles W-Z, which are not in the longitudinal set, but which have been cited by documents within the set. In this example we assume that the time window of the longitudinal data set is short (e.g., 5 years), and thus equate the research front with this short window of current articles. Articles A-E might be published in the last 2 years, articles M-P would be published 3-5 years ago, and articles W-Z would be published more than 5 years ago.



**Figure 1. Example of how different citation-based mapping approaches partition the same set of documents. The gray box represents the documents within a dataset. Documents W, X, Y and Z are documents outside the set, but are referenced by documents within the set. Solid arrows represent citations within the set. Dashed arrows represent citations to documents outside the set. Ovals in each panel show how the documents might be clustered by each approach.**

We first consider clusters of articles formed within the longitudinal data set or research front if only references within the set are considered. Direct citation, where articles are linked if one references another, only considers links from within the set. In the example of Figure 1, cluster (A,M,N) will form because A cites both M and N; cluster (C,D,O) will form because C and D both cite O, and cluster (E,P) will form from the link between E and P. Article B will not be clustered because it doesn't cite any other article within the set. Bibliographic coupling links documents that reference the same set of cited documents. In the case of Figure 1 only one cluster will form if only internal links are considered; cluster (C,D) forms since C and D both cite O. The remaining documents are not placed in clusters since none of them cite a reference paper that is cited by any other document within the set.

The word *co-citation* is used in different science mapping studies to denote two different processes. Here we differentiate between co-citation clustering, which is simply the formation of clusters of co-cited documents, and co-citation analysis, which takes the result of co-citation clustering, and then assigns current papers (or papers from the research front) to the co-citation clusters. Of the two, co-citation analysis is the more standard approach. In the case of Figure 1, if only within-set links are used, co-citation clustering will form a single cluster (M,N) since both M and N were cited by A. If the process is expanded to co-citation analysis, the cluster will expand to include paper A, since A cites M and N.

This example shows some of the broader effects associated with the different citation-based approaches. In a longitudinal data set where links are restricted to those within the set, bibliographic coupling is able to cluster very recent papers but clusters fewer of the very old papers, while co-citation clustering does the opposite – it clusters the older papers, but cannot cluster the most recent papers that have not yet been cited. Direct citation clusters documents more evenly across the time window, and tends to cluster a larger number of documents than either bibliographic coupling or co-citation processes.

Figure 1 also shows the clusters that will result using the different citation-based approaches if external references are used in addition to the within-set links. For bibliographic coupling, the inclusion of external references greatly increases the number of papers in the research front that can be clustered. In the case shown here two clusters will form: cluster (A,B,C,M,N) because these documents cite W and X, and cluster (D,E,O,P) from citing of Y and Z. Article C cites references from both clusters, but would likely end up in the (A,B...) cluster because it cites two core references that form that cluster, while only citing one reference that forms the other cluster.

Co-citation clustering approaches also benefit greatly from the use of external references, because the external references can be members of co-citation clusters. In the Figure 1 example, co-citation clustering using external references would form two clusters: (M,N,W,X) and (O,P,Y,Z). One of the features of co-citation analysis is the possibility of fractionally assigning papers in the research front to multiple clusters. In the example shown, assignment of the research front papers to the two co-citation clusters would add papers A and B to the first cluster, and papers D and E to the second cluster. Paper C could be split between the two clusters in fractions of 2/3 and 1/3, respectively, since it has 2 links to the first co-citation cluster and one link to the second cluster. Note that both the bibliographic coupling and co-citation analysis

approaches based on both internal and external references cluster more documents than does the direct citation approach.

The example of Figure 1 is obviously greatly simplified from actual citation networks. Nevertheless, it does illustrate the differences between approaches, and also shows that, even in this highly simplified case, there are enormous differences between the resulting cluster solutions, thus justifying the need for a study to compare the accuracies of different cluster solutions. It also shows that the inclusion of external references is highly beneficial from a coverage standpoint. If the application for a particular literature map requires full coverage (or nearly full coverage) of that literature, within-set linkages are not sufficient.

For completeness, we also note here that the direct citation approach does have a variant (not shown in Figure 1) that we have never seen published. In the same way that co-citation analysis fractionally assigns papers from the research front to co-citation clusters of reference papers, one can consider the direct citation clusters as clusters of reference papers, and then fractionally re-assign the papers to clusters based on the reference structure. To do this, one must make the assumption that each paper cites itself as well as the other papers in the reference list. This method does not increase the number of papers that are assignable, but does give the option of splitting the papers between multiple clusters if that feature is desired.

### ***Accuracy studies***

Although quantitative comparison of the accuracies of maps generated from different similarity approaches is a relatively recent topic, researchers have, from the beginning, sought to show that their maps are accurate in the sense that they correspond with reality. One early example of this is a map that shows not only disciplinary areas, but also labels the linkages between those disciplinary areas with the topics that are the main sources of those linkages (Small & Garfield, 1985). Showing detail that corresponds to our perception of reality is one way of establishing the face validity of a map of science.

Braam, Moed & van Raan (1991) took the next logical step and, in addition to giving a most open and honest assessment of co-citation analysis from the points of view of both protagonists and detractors, calculated the within-cluster and between-cluster coherence of content words (indexing terms and classification codes) associated with co-citation clusters. They concluded that co-citation analysis does generate clusters that are topically coherent, and that correspond loosely with research specialties. This association is not prescriptive, or one-to-one, but rather several clusters could be associated with a single specialty. Nevertheless, the study did establish quantitatively that co-citation analysis produces research fronts (recent papers assigned to co-citation clusters) that are topically coherent and relatively distinct from each other.

Recent studies that compare the cluster solutions from different similarity approaches have been done at two different levels – journals and documents. Two different research teams have compared the accuracies of similarity approaches in journal mapping. The current authors used a set of 7,121 journals from the 2002 combined SCIE/SSCI indexes and compared various intercitation and co-citation based similarity measures (Boyack et al., 2005; Klavans & Boyack, 2006a). Their maps were compared using mutual information and pairwise binary overlap

metrics with the ISI subject categories as the standard of comparison. Their primary finding regarding similarity measures was that normalized measures (e.g., cosine, Jaccard, Pearson correlation) generate far more accurate maps than those based on raw citation counts. More recently, researchers at KU Leuven have compared a variety of citation-based, text-based, and hybrid similarity approaches using a set of 8,305 journals from the 2002-2006 Web of Science (Janssens et al., 2009; Liu et al., 2010). The most recent of these studies (Liu et al., 2010) compared 5 citation-based approaches (including cross-citation<sup>1</sup>, co-citation, and bibliographic coupling), 5 text-based approaches (including TFIDF and LSI-TFIDF), and 9 different weighting schemes to combine the matrices of the 10 text- and citation-based models. The 19 models were each partitioned into 22 clusters and then compared using an adjusted Rand index and the normalized mutual information (NMI) metric with the 22 high-level ESI (Essential Science Indicators) categories as the basis of comparison. The WEAC-AL (weighted evidence accumulation) hybrid weighting scheme performed best overall, with NMI values 6.6% and 3.8% higher than the best text- and citation-based approaches, respectively. Interestingly, of the pure text- and citation-based approaches, the binary approach performed best in each case. This is a somewhat counterintuitive result in that one might expect continuous measures that give fine-grained differentiation in similarities to generate a more accurate cluster solution than a simple (0,1) binary measure, and deserves more investigation.

Comparisons at the document level started to appear at about the same time as the earliest journal level study mentioned above. The KU Leuven group, in addition to their work with journals, also investigated similarity approaches at the document level (Janssens et al., 2006). They partitioned 5,188 bioinformatics articles into 2 and 7 cluster solutions using 13 different similarity approaches, including TFIDF, LSI, bibliographic coupling and two variants, and many different hybrid approaches that combined the text and citation vectors in various ways. Cluster qualities of the 13 solution sets were compared using Silhouette values calculated from the MeSH term distributions for each cluster. Using this metric, the hybrid approaches outperformed both text-only and citation-only approaches. The KU Leuven group also compared the accuracies of text-based approaches using full text articles as opposed to using just the words from titles and abstracts for clustering (Glenisson, Glänzel, Janssens et al., 2005; Glenisson, Glänzel, & Persson, 2005), and found that full text outperformed titles and abstracts on sets of 19 and 85 documents from the journal *Scientometrics*.

Cao & Gao (2005) examined a set of 4,330 articles in the machine learning area, and found that adding citation information to their textual feature vectors resulted in a 3-4% gain in classification accuracy when compared to the known classifications for the documents. They also found that feature vectors that included 2-word phrases in addition to single words gave slightly higher accuracies than feature vectors containing only single words.

Jarneving (2005) compared the representations of the research front as calculated using bibliographic coupling and co-citation analysis. Using a dataset comprised of over 73,000 articles, a 10% sample was taken and then further reduced using coupling strengths and cluster size thresholds, to ultimately compare a bibliographic coupling set of 1,691 articles in 88 clusters with a co-citation set of 2,094 articles in 96 clusters. The overlap between the two sets was only

---

<sup>1</sup> Cross-citation as defined by Liu et al. (2010) is identical to intercitation as defined by Boyack et al. (2005). Both studies generated counts at the paper level, aggregated to journals, and ignored citation direction between journals.

612 articles. Although there was a clear conclusion that the two methods produced very different representations of the research front (using analysis of words patterns associated with the two cluster solutions), no conclusion could be arrived at as to which method produced the most accurate representation of the research front.

Klavans & Boyack (2006b) generated 4 bibliographic coupling and 4 co-citation cluster maps of a much larger document space, the 2002 SCIE/SSCI set of documents. The bibliographic coupling maps contained 731,000 articles, while the co-citation cluster maps contained 719,000 cited references. Although maps from the two approaches were not compared, this study showed that normalized measures (e.g. cosines) produced more accurate maps than raw count-based measures, and also that cluster solutions tending toward smaller clusters were more accurate than cluster solutions tending toward larger clusters.

Calado et al. (2006) clustered two separate pre-classified collections of web documents (over 40,000 documents each) using five different similarity approaches, including bibliographic coupling, co-citation clustering, Amsler (a linear combination of bibliographic coupling and co-citation clustering), and text-based TFIDF. Using the F1 (combined precision-recall) metric, they found that the text-based approach was far better than any of the citation-based approaches if only internal (or within-set) links were used, while the co-citation and Amsler approaches did far better than the text-based approach if both internal and external links were used. The same research team ran a similar study using the ACM8 document collection, a set of 6,680 documents from 8 first-level categories in the ACM digital library (Couto et al., 2006). However, the results were different in this case. Using the external links that were available (within the ACM full set) along with internal links, the Amsler similarity approach had the best performance, followed closely by bibliographic coupling. These citation-based approaches did better than the TFIDF cosine approach by 3-10% depending on clustering method. Co-citation performed worst in this study; this poorer performance was likely due to the lack of availability of external citations to these ACM documents from articles outside the ACM collection, and points out the advantage of working with comprehensive data when doing accuracy studies.

Shibata et al. (2009) compared cluster solutions from direct citation, bibliographic coupling, and co-citation clustering on three data sets ranging in size from 3,510 to 23,459 articles, using within-set links only. Using measurements based on cluster size, citation speed, and linkage density, they found that direct citation was quickest and best at detecting emerging research fronts while co-citation was worst. These results can be directly correlated with the effects of using within-set links mentioned in the discussion around Figure 1. Given that co-citation clustering using only within-set links will not cluster any recent paper that has not yet been cited within the set, this technique biases against recent papers, and biases against detection of the research front. Co-citation analysis would certainly have been a better choice for comparison in this study than was co-citation clustering.

Finally, Ahlgren and co-workers tested both 1<sup>st</sup> and 2<sup>nd</sup> order similarity approaches on a set of 43 documents from the *Information Retrieval* journal (Ahlgren & Colliander, 2009a; Ahlgren & Jarneving, 2008). Comparing the results against a by-hand classification of these documents into 15 clusters with a Rand index, they found that bibliographic coupling did very poorly, while text and hybrid approaches did much better. Among 1<sup>st</sup> order similarity approaches a hybrid using a

linear combination of TFIDF-SVD and bibliographic coupling approaches was best, while among 2<sup>nd</sup> order approaches the original TFIDF was best. Second order approaches performed better than their associated first order approaches in all cases, and as such are very promising. Ahlgren and Colliander (2009b) followed this up with a study of another data set containing 58 documents related to science metrics and found that bibliographic coupling outperformed TFIDF using a variety of clustering techniques. We question whether the results from these studies (and those of Glenisson et al. mentioned earlier) with such small samples and conflicting outcomes can be generalized or scaled in any meaningful way.

From the foregoing discussion of similarity metric comparisons, although the data are somewhat sparse, there seems to be a growing body of results suggesting that properly constructed hybrid text-citation approaches can lead to more accurate maps of science than can be generated from text-only or citation-only approaches. In this article, we do not address text-only similarity approaches, but we will address the difference between citation-only similarity approaches and a simply hybrid approach that incorporates textual information into a citation-based approach.

Regarding accuracy, there is one additional observation we would like to make. Despite the relatively small number of quantitative accuracy studies that have been published, most current science mapping studies do attempt to consider accuracy in a qualitative fashion. Most mapping studies, whether based on documents, authors, or journals, tell stories about or explain some of the associations observed in their maps as a means of self-validation. By saying this we do not mean to suggest that this self-validation lacks meaning. In fact the opposite is true – the observations and stories associated with the partitioning, structure, and dynamics of these maps, and our association of these observations with reality, are the things that give our maps face validity, make them compelling, and make us want to dig a little further. The stories and the potentially actionable results are what drive us to seek a more accurate map.

## **Data and Methods**

### *Study corpus*

The purpose of the full study that was performed for NIH was to find the most accurate science mapping solution on a very large corpus of biomedical literature – in essence, to determine how to most accurately map all of the medical sciences literature. As mentioned previously, even though we only report on citation-based and hybrid approaches here, the full study compared text-based, citation-based, and hybrid text-citation approaches. A study corpus was needed that would be large enough to provide definitive results, and that would not be biased toward either text- or citation-based approaches. We thus generated a corpus for which we had both sufficient text and citation information for all articles.

Given that NIH hosts the MEDLINE database, its contents are an implicit definition of what the NIH considers to be medical science. We thus felt it best to base our study corpus on MEDLINE, and to add citation data as needed. Scopus records were matched to MEDLINE records to generate a set of records with both textual and citation information. This set of records was then limited to those articles published from 2004-2008 that contained abstracts, at least 5 MeSH

terms, and at least 5 references in their bibliographies, resulting in a study corpus of 2,153,769 unique scientific articles. The process used to generate this corpus was as follows:

- 1) Records from MEDLINE were matched to Scopus records to generate a one-to-one matching between records, to identify those records (articles, etc.) for which MeSH terms, titles, abstracts, and references are available. Matching was carried out on a segment of both databases containing publications from 2003 to 2008 to ensure that at least 2 million records would be matched, using the following process:
  - a. Over the past several years, we have maintained a matched list of MEDLINE and Scopus journals. This list was used to add the Scopus journal ID number to 99.8% of the MEDLINE records (corresponding to 7611 different MEDLINE journal abbreviations) from 2003 to 2008.
  - b. A sequence of steps using different criteria was then used to match MEDLINE and Scopus article data. Matching was done without replacement; if a particular MEDLINE record was matched in one step, it was removed from the list and not available for matching in a subsequent step. The matching criteria, in order, were:
    - i. Journal ID AND starting page AND (volume OR pubyear) AND soundex<sup>2</sup>(title)
    - ii. Journal ID AND volume AND soundex(title)
    - iii. Journal ID AND pubyear AND soundex(title)
    - iv. Journal ID AND soundex(title)
    - v. Given that each of the above matching steps (1-4) generated some duplicate matches to PMIDs (PubMed ID), the matched set was restricted to unique PMID-ScopusID matches (meaning each PMID and ScopusID could only appear once in the full list).
    - vi. For any duplicate matches, matches where the first five initials of the first author's last name did not match were removed.
    - vii. All remaining unmatched PMID were left as unmatched. Results of these matching steps are given in Table 1.

As shown in Table 1, the overall matching rate (unique PMID to ScopusID) for the entire set of MEDLINE documents from 2003 to 2008 was 95.3%. The matching rate for 2008 (92%) is lower than for previous years (over 96%) because the full 2008 data were not yet available in the Scopus raw data that we were using.

**Table 1. Efficiency of matching Scopus to MEDLINE records.**

Step	Counts	Fraction	Total matched
0 – initial MEDLINE records	3,647,481		
1 – matching criteria i	2,847,197	78.06%	2,847,197
2 – matching criteria ii	557,991	15.30%	3,405,188
3 – matching criteria iii	91,985	2.52%	3,497,173
4 – matching criteria iv	2,676	0.07%	3,499,849
5-7 – remove duplicate/false matches (v-vii)	3,475,573	95.29%	3,475,573

<sup>2</sup> “Soundex” is a function in MySQL that strips all non-alphanumeric characters from text, and converts the remaining text to a string based on phonetics. Two strings that sound the same, but that have different spellings, can thus have the same soundex value. Use of the soundex function allows us to match some records where there are simple misspellings or punctuation differences in the article titles between the two databases.

- 2) Additional data were added to the matched data: numbers of references from Scopus, numbers of MeSH terms from MEDLINE, and the existence of a MEDLINE abstract. These data were then used to limit the set of records to those with sufficient text and citation information to form an appropriate corpus for this study. It was clear from the numbers that all 6 years (2003 to 2008) of records were not needed to give a set of around two million documents. We thus restricted the set to five years (2004 to 2008). Numbers of documents by year, using different limitations, are given in Table 2.

**Table 2. Numbers of documents by year using different limitations.**

Year	MEDLINE	In Scopus	A	R	A or R	A and R	Final
2004	575,938	553,743	454,023	436,421	489,545	400,899	389,353
2005	603,166	579,359	480,477	469,777	517,773	432,481	420,059
2006	626,895	605,734	504,747	498,328	547,663	455,412	442,743
2007	644,457	620,386	523,805	520,196	566,781	477,220	464,479
2008	650,069	597,839	506,852	490,034	547,110	449,776	437,135
Total	3,100,525	2,957,061	2,469,504	2,414,756	2,668,872	2,215,788	2,153,769

A – has an abstract; R – has  $\geq 5$  references

It is interesting to examine the numbers from Table 2. Only 81.7% of the MEDLINE records in Scopus have 5 or more references, only 83.5% of the records have abstracts, and only 74.9% of records have both. This suggests that if one wants a map with full coverage, a hybrid approach would be necessary – otherwise, a citation-based map would be missing around 18% of documents, and a text-based map would be missing 16%.

For the study corpus, it was decided to keep those documents with an abstract, at least five references, and at least 5 MeSH terms. In addition, there were several hundred articles with very large numbers of references. In our experience, articles with large numbers of references can lead to over-aggregation of citation clusters. Thus, we arbitrarily set a threshold of 400 references; article with more references than this were excluded from the corpus. The final numbers of articles by year that met these criteria are listed in the final column of Table 2.

The accuracy of our matching process was checked by accessing the PMIDs as indexed in the Scopus raw data. Of our 2,153,769 ScopusID to PMID matches, we found that the Scopus raw data did not contain a PMID for 129,300 (6.0%) of the ScopusIDs. Spot-checking of the corresponding records from each database showed that our matches were indeed valid matches. Within the 2,024,469 records for which Scopus had a PMID, our matched PMIDs were only different in 27 cases. The matching process used above thus has an accuracy rate of  $> 99.99\%$ , and use of our matching process provided a more complete corpus for this project than if Scopus PMIDs alone had been used to link the reference data from Scopus to the MEDLINE documents.

### *Similarity approaches*

Three different citation-based maps were generated using the reference information associated with our corpus of 2,153,769 documents. Each method requires different processing from a single starting point – the full list of citing:cited document ID pairs from the corpus. This initial list consists of 80,754,581 citing:cited pairs that reference 15,503,380 unique reference papers.

Co-citation analysis (CCA) method: The general process for co-citation analysis is to 1) identify a set of reference papers, 2) calculate the similarity between pairs of reference papers using co-citation counts, 3) calculate co-citation clusters of reference papers using the similarity values, and 4) fractionally assign the current (or research front) papers to the co-citation clusters based on location of their references. The first two steps, identification of the reference set and calculation of the similarity values, are detailed here. The clustering step will be detailed later.

Reference papers were filtered using the following formula which is based on the standard co-citation practice at SciTech Strategies:

- The articles from the corpus were grouped by publication year and separate citing:cited pair files were generated for each of the five publication years.
- For each yearly set, the citations to each unique reference paper were counted.
- For each yearly set, reference papers that met the following criteria were retained in the set:  
 $(\text{age} = 0 \text{ and } \text{ncited} \geq 3) \text{ OR } (\text{age} < 3 \text{ and } \text{ncited} \geq (\text{age}+1)) \text{ OR } \text{ncited} \geq 5$   
 where age = citing publication year – cited publication year.
- The retained references from the five yearly sets were combined to form the full set of references, resulting in a set of 2,473,611 unique references. The original citing:cited pair list was then filtered to include only those pairs where the cited document was in the set of 2,473,611 references, resulting in a pairs file containing 50,221,140 citing:cited pairs. All references were cited at least 4 times over the five year period (the criteria in the third bullet was by year), and the most highly cited reference was cited by 25,579 (1.1876%) of the citing documents.

Co-citation similarities were calculated as:

- Co-citation frequencies,  $C_{i,j}$ , between pairs of reference documents  $i$  and  $j$  were calculated from the citing:cited pairs list.
- Each co-citation frequency was modified using

$$F_{i,j} = 1/\log(p(C_{i,j}+1)) \text{ where } p(C_{i,j}+1) = C_{i,j} (C_{i,j}+1)/2. \quad (1)$$

- K50 (modified cosine) values were calculated from each  $F_{i,j}$  value as:

$$K50_{i,j} = K50_{j,i} = \max \left[ \frac{(F_{i,j} - E_{i,j})}{\sqrt{S_i S_j}}, \frac{(F_{j,i} - E_{j,i})}{\sqrt{S_i S_j}} \right] \quad (2)$$

where  $E_{i,j} = \frac{S_i S_j}{(SS - S_i)}$ ,  $S_i = \sum_{j=1}^n F_{i,j}$ ,  $j \neq i$ ,  $SS = \sum_{i=1}^n S_i$ .

$E$  is an expected value of  $F$ , and varies with  $S_j$ ; K50 differs from most other measures in that it is a relative measure that subtracts out the expected value. Thus K50 will only be positive for those reference paper interactions that are larger than expected given the matrix row and column sums. Note also that although  $E_{i,j} \neq E_{j,i}$ , the differences in these values are typically too small to be of consequence for paper-level similarities, even though they can be quite large for journal-level similarities.

The full K50 matrix is too large for our clustering routines, thus we filter the similarities to generate a reduced size similarity file. Filtering was done by 1) removing all pairs with negative

K50 values, 2) sorting the remaining list by reference and descending K50 value, 3) removing all references which had more than 15 K50 values tied in the first position (because they were obviously non-differentiating), and 4) using the total degree distribution for each reference, and scaling the  $\log(\text{degree})$  values to a 5-15 scale. The degree for each reference thus determines how many pairs that reference brings into the final similarity file, varying between 5 and 15. We call this a top-n similarity file. Although our clustering routines ignore (A:B – B:A) duplicates, it is useful to de-duplicate the similarity file for efficiency. After de-duplication, the total number of cited document pairs in the cited document similarity file was 15,537,317.

*Bibliographic coupling (BC) method:* The general process for bibliographic coupling is to 1) identify a set of recent papers, 2) calculate the similarity between pairs of papers using bibliographic coupling counts, and 3) assign citing papers to clusters using the similarity values. The first two of these steps are detailed here:

While we did need to filter the reference papers for the co-citation calculation, no filtering of citing documents was needed because the full set of 2.15 million documents is well within the range of what our clustering routines can manage. The full citing:cited pair list was used to generate a similarity using the following process:

- Reference papers cited more than 500 times within the set were removed from the citing:cited pair list to avoid the over-aggregation associated with coupling based on highly cited references.
- Bibliographic coupling frequencies,  $B_{i,j}$ , between pairs of citing documents were calculated from the filtered citing:cited pair list. The total number of citing document pairs (full matrix) with a non-zero coupling count was 170,835,050.
- Bibliographic coupling K50 values were calculated from frequencies  $B_{i,j}$ , using the same method and equations listed above for co-citation analysis, with values  $B$  replacing values  $C$ . The full list of similarities was filtered to a top-n similarity file as detailed above. After de-duplication, the total number of citing document pairs in the bibliographic coupling similarity file was 14,159,303.

*Direct citation (DC) method:* The general process for direct citation is to 1) identify the similarity pairs within the data set, 2) calculate similarities between pairs of papers, and 3) assign papers to clusters using the similarity values. The first two of these steps are detailed here.

For direct citation, since all references pairs are within the set, no pre-filtering is necessary. However, documents that either do not cite or are not cited by any other document within the set will not be a part of the calculation. The within-set citing:cited pairs for this calculation were determined by finding all pairs where both the citing and cited document were within the set. This reduced the citing:cited pair list to 23,218,091 pairs. This list was used to generate a similarity using the following process:

- Each citing:cited pair was assigned a weight  $wt = 1/n$  where  $n$  is the total number of papers cited by the citing paper in the pair. Thus, each citing paper contributes a total weight of 1.0 to the initial direct citation counts. We did not see the need to use a more complex weighting system that would account for times cited as well as number of references.

- Since this citing:cited:wt list is directional, and thus comprises only the upper half of a full citation matrix, this list was flipped (cited:citing:wt) and concatenated to the upper half, thus forming a full, symmetric matrix of fractional direct citation counts,  $D_{ij}$ .
- The next two steps were the same as the final two steps listed for the co-citation method above: calculating K50 values using frequencies  $D$  in place of modified frequencies  $F$ , and filtering the full list of similarities to a top-n similarity file. After de-duplication, the total number of citing document pairs in the direct citation similarity file was 7,581,738.

*Hybrid similarity (HYB) method:* A sample hybrid similarity approach using both references and words was calculated to provide a proof of concept test as to whether a text-citation hybrid similarity approach might be comparable, or perhaps even better, in terms of performance, than the best text-based and best citation-based similarity approaches. Our hybrid test was done after all of the work using all other similarity approaches was completed, and thus was designed to be simple, and to take advantage of the lessons learned from working with the results from the citation-based similarity approaches.

The hybrid similarity method used here was identical to the bibliographic coupling method detailed above with the difference being that the coupling was done on both references and words from the title/abstract matrix. Citing paper:word pairs were added to the citing:cited paper list, and words were treated as if they were references. Only words occurring in between 4 and 500 documents within the corpus were added. These thresholds were chosen to match the thresholds used on reference papers in the bibliographic coupling calculation. The process detailed above for bibliographic coupling was then used on this hybrid text-citation matrix.

### *Clustering*

Similarity files from each of the similarity approaches above were run through a standardized and very robust clustering process to generate sets of document clusters. The same clustering method was used for all similarity approaches; thus the clustering method should not contribute to any variability in the final results. The clustering process was comprised of four main steps:

- 1) The DrL<sup>3</sup> (formerly VxOrd) graph layout routine (Martin, Brown, Klavans, & Boyack, 2008) was run using a similarity file as input, and using a cutting parameter of 0.975 (maximum cutting). DrL uses a random walk routine and prunes edges based on degree and edge distance; long edges between nodes of high degree are preferentially cut. A typical DrL run using an input file of 2M articles and 15M edges will cut approximately 60% of the input edges, where an edge represents a single document-document similarity pair from the original similarity file. At the end of the layout calculation, each article has an x,y position, and roughly 40% of the original edges remain.
- 2) Papers were assigned to clusters using an average-linkage clustering algorithm (Klavans & Boyack, 2006b). The average-linkage clustering algorithm uses the article positions (x,y) and remaining edges to assign papers to clusters. Once the clusters are generated, the full list of pairs of papers that co-occur in a cluster are generated for each solution. For example, if papers A, B, C, and D are in a cluster together, the set of pairs will be AB, AC, AD, BC, BC, and CD.

---

<sup>3</sup> Sandia National Laboratories has recently renamed DrL to OpenOrd, which is freely available at <http://www.cs.sandia.gov/~smartin/software.html>.

- 3) Steps (1-2) were run 10 separate times using 10 different random starting seeds for DrL, and thus giving rise to 10 unique cluster solutions for the same similarity file. Different starting seeds (i.e., different starting points for the random walk) will give rise to different graph layouts and different (but typically highly overlapping) sets of remaining edges. We use these differences to our advantage in this clustering process.
- 4) Those paper pairs that appear in 6 or more out of the 10 DrL solutions are considered to be the robust pairs, and are listed in a separate file.<sup>4</sup> This list of pairs is then used as the input edges to the same average-linkage clustering algorithm used in the previous step. Using this input, the algorithm essentially finds and outputs all distinct graph components. Each separate component is a cluster, and these clusters are referred to as level 0 clusters.
- 5) Logic dictates that a cluster should have a minimum size; otherwise there is not enough content to differentiate it from other clusters. In our experience, a cluster should contain a minimum of approximately five papers per year (or 25 papers over the five year length of the corpus) to be considered topical.<sup>5</sup> Thus we take all clusters with fewer than 25 papers, and aggregate them. This is done by calculating K50 similarities between all pairs of level 0 clusters, and then aggregating each small cluster (< 25 papers) with the cluster to which it has the largest K50 similarity until no clusters with < 25 papers remain. K50 values are calculated from aggregated modified frequency values (the  $1/\log(p(C+1))$  values) where available, and from the aggregated top-n similarity values in all other cases. The resulting aggregated clusters are known as level 1 clusters.

Previous experience has shown that a cluster solution based on the combination of 10 DrL runs is much more robust than that from a single DrL run. For example, using a co-citation model of roughly 2.1M documents and 16M edges, the adjusted Rand index<sup>6</sup> between pairs of single DrL solutions was 0.32, while the adjusted Rand index between pairs of 10xDrL solutions was over 0.80. Requiring that papers be paired in the same cluster in 6 out of 10 separate DrL solutions thus limits the final solution to only those pairs and sets of pairs (and thus the resulting clusters) that are relatively robust. However, this robustness can also have a deleterious effect on the final coverage of a solution – any papers that are not paired with another particular paper in at least 6 out of the 10 DrL runs will drop out of the solution. To some degree, this is in itself a measure of the robustness (or conversely, ambiguity) in the similarity approach. Solutions in which many papers are dropped do so because of ambiguity in the overall similarity space of the document set. Thus, similarity approaches that generate solutions dropping many nodes can be thought of as more ambiguous than similarity approaches that drop few nodes.

---

<sup>4</sup> The 6/10 criteria was arrived at through testing. A criteria of 7/10 solutions leads to insufficient coverage, while a criteria of 5/10 solutions leads to cluster chaining and the formation of a giant component. Both of these conditions are undesirable. The 6/10 criteria thus represents a solution with sufficient coverage and a reasonable cluster size distribution.

<sup>5</sup> This is based on an old, undocumented assumption among many bibliometricians that there must be a critical mass of around 5 papers per year in a cluster for the cluster to represent a specific and measurable topic in science. We have chosen to aggregate small clusters to this level, but others may choose to handle small clusters in a different manner. There is no consensus in the community as to how the issue of small clusters should be dealt with.

<sup>6</sup> The Rand index is an overlap measure based on the partitioning of paired elements in two data sets, see [http://en.wikipedia.org/wiki/Rand\\_index](http://en.wikipedia.org/wiki/Rand_index). The adjusted Rand index adjusts the Rand index for chance, and is a more stringent test than the Rand index.

### *Assignment of reference front articles to co-citation clusters*

For the bibliographic coupling, direct citation, and hybrid approaches, the results of the clustering step are final because it is the papers from the corpus themselves that are being clustered. However, one step remains for the co-citation analysis approach – the papers from the corpus must be assigned to the co-citation clusters using their reference lists. This step is done at the level 0 cluster level (between clustering steps 3 and 4 above), and before aggregation to level 1 clusters, as follows:

- For each citing paper the number of references to papers in each of the level 0 clusters is calculated from the citing:cited pairs list.
- For each citing paper, the maximum number of references to level 0 clusters was calculated, and the paper was labeled as *unambiguous* if that maximum number was greater than 1. Citing papers with a maximum number of 1 were labeled as *ambiguous*.
- A *t-value* is calculated for each citing paper, level 0 cluster pair where *t* is the number of references to papers to the level 0 cluster divided by the square root of the number of papers in the level 0 cluster. Thus, the *t-value* is related to the fraction of the cluster that is cited by the citing paper. For unambiguous papers, *t* is only calculated for clusters where the number of references to the cluster is greater than 1. This avoids long tails for unambiguous papers.
- For each citing paper, t-values are normalized to sum to 1.0. These normalized t-values are used as the current paper to level 0 cluster weights.
- For each citing paper, if it is also a reference paper in a level 0 cluster, the fractions are adjusted; each fraction is set to one half of its previous value, thus giving each a summed fractional value of 0.5. The remaining 0.5 fraction is assigned to the level 0 cluster to which the paper belongs as a reference paper. The citing paper, level 0 fractions are then re-summed to give the final fractional assignments.

Once the citing paper assignments have been made, then clustering step (4) above is applied to merge clusters that contain fewer than 25 papers (using summed fractional counts) into level 1 clusters.

### *Clustering results*

Metrics from the 10xDrL cluster solutions from each of the similarity approaches are given in Table 3. Cluster size distributions for the cluster solutions are shown in Figure 2.

**Table 3. Clustering characteristics of the different cluster solutions.**

Method	# Articles covered	% Coverage	# Level0 Clusters	# Level1 Clusters	Level1 Max Size
CCA	2,118,644	98.37%	188,561	32,184	3245
BC	2,081,022	96.62%	207,764	32,782	778
DC	1,996,050	92.68%	456,112	50,719	376
HYB	2,085,577	96.83%	198,122	31,121	596

The clustering results lead to several observations. First, the direct citation method gives by far the largest number of clusters, and has the smallest cluster sizes. In fact, there were nearly 10,000 level 0 clusters with fewer than 25 members (including 7,387 with only two members) that could not be aggregated into level 1 clusters because there was simply no direct citation relationship

with any member of any other cluster. Thus, these level 0 clusters were carried forward to be level 1 clusters. In addition, only 92.7% of the corpus was placed into clusters by the direct citation method. These results related to direct citation are not surprising – they correlate very well with the observations associated with the example in Figure 1.

Second, co-citation gives by far the largest cluster sizes for its largest clusters. However, once the few clusters in the co-citation solution of size > 600 articles are accounted for, the cluster size distributions for co-citation analysis, bibliographic coupling, and the hybrid approach are very similar. All three of these approaches have coverages of greater than 96.6%, with co-citation leading at 98.4%. All three of these approaches provide very high coverage of the corpus using the clustering parameters selected. The similar cluster size distributions and coverages also suggest that these factors will not negatively impact the accuracy comparisons to be shown later.

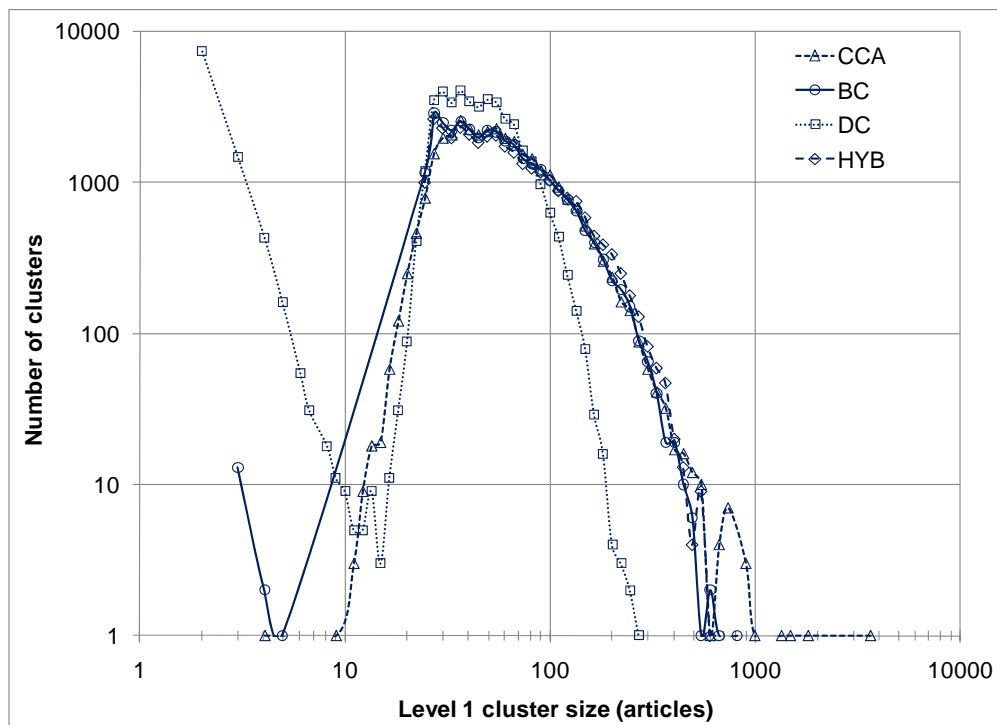


Figure 2. Level 1 cluster size distributions.

## Accuracy Metrics

Given that the main purpose of this study was to determine which similarity approach would create the most accurate map of the medical sciences, we needed a way to measure the relative accuracies of each cluster solution. We settled on two measures – the first measures the within-cluster textual coherence, and the second is a new measure designed to show which similarity approach best concentrates grant-to-article linkages. We note that the notion of a research front includes both topical and social components, and further note that both of these measures focus on the topical content rather than the social aspects of the document clusters. We are not aware of an accuracy measure based on the social component of document clusters; such a measure would be a welcome addition to the literature.

### **Textual coherence**

The quantity that is used here to measure textual coherence is the Jensen-Shannon (JSD) divergence (Lin, 1991). It is used to quantify the distance (or divergence) between two (or more) probability distributions. JSD<sup>7</sup> is calculated for each document from the word probability vector for that document, and from the word probability vector for the cluster in which the document resides as:

$$JSD(p, q) = \frac{1}{2} D_{KL}(p, m) + \frac{1}{2} D_{KL}(q, m) \quad (3)$$

$$\text{where } m = (p+q)/2 \text{ and } D_{KL}(p, m) = \sum (p_i \log (p_i/m_i))$$

and  $p$  is the frequency of a word in a document,  $q$  is the frequency of the same word in the cluster of documents, and  $D$  is the well-known Kullback-Leibler divergence. JSD is calculated for each cluster as the average JSD value over all documents in the cluster.

JSD is a divergence measure, meaning that if the documents in a cluster are very different from each other, using different sets of words, the JSD value will be very high, or close to 1.0. Clusters of documents with similar sets of words – a less diverse set of words – will have a lower divergence. The use of JSD is not limited to sets of words, but is commonly used in mathematical statistics and statistical physics applications (Grosse et al., 2002), and more recently in bioinformatics (Sims, Jun, Wu, & Kim, 2009).

JSD varies with cluster size. For example, a cluster with 10 very different documents will have a larger set of unique elements, and thus a higher divergence value than a cluster with only 3 very different documents. The maximum possible JSD values for various cluster sizes will occur when the documents in the cluster have completely different sets of elements. These maximum divergence clusters can be approximated, for a particular corpus, by forming random clusters of documents from that corpus. We have calculated JSD values for randomly formed clusters of different sizes from the study corpus, as shown in Figure 3. Each measured divergence value in Figure 3 is an average of the divergence values from a very large number of random clusters (e.g., 5000 random clusters of size 20, 5000 random clusters of size 100, 1000 random clusters of size 500). A curve fit of the measured values was used to estimate the JSD values for every cluster size from 2 to 1000. The very small error bars (one standard deviation) on the curve in Figure 3 show that the random divergence values have a very small variance, and suggest that JSD is inherently not a noisy measure.

Coherence is calculated from divergence values for each cluster  $i$  as:

$$Coh_i = JSD(rand)_i - JSD(actual)_i \quad (4)$$

where  $JSD(rand)$  is the random divergence for the particular cluster size. The average coherence value for an entire cluster solution is then calculated as a weighted average:

$$Coh = \sum n_i * Coh_i / \sum n_i \quad (5)$$

---

<sup>7</sup> We use the simplified JSD formulation for two distributions of equal weights used in Sims et al. (2009), also found on Wikipedia ([http://en.wikipedia.org/wiki/Jensen-Shannon\\_divergence](http://en.wikipedia.org/wiki/Jensen-Shannon_divergence)).

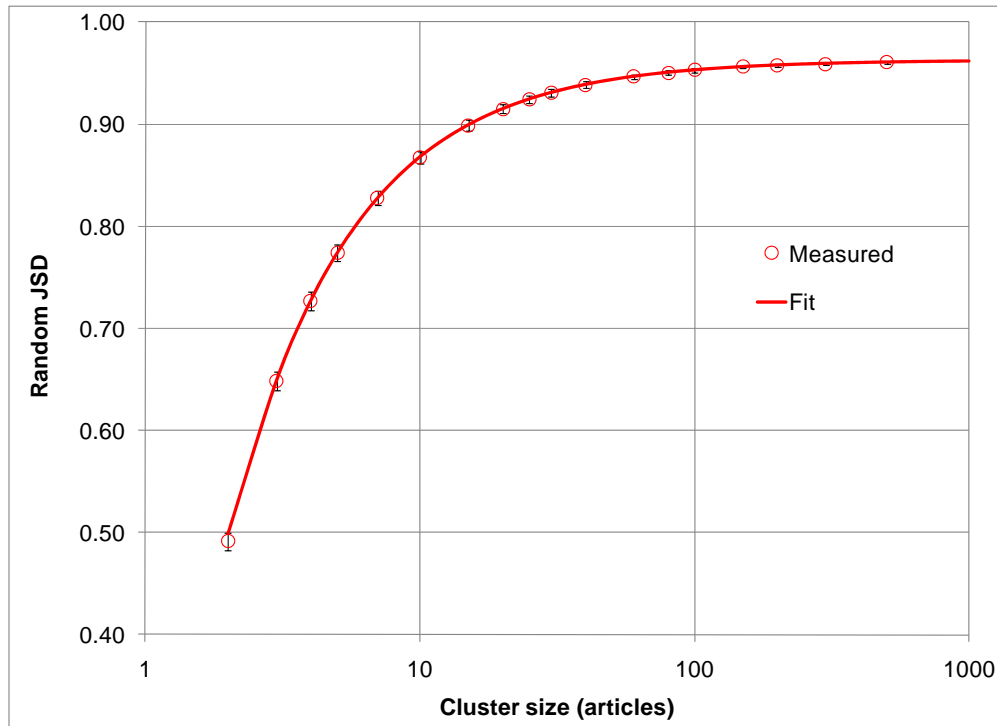


Figure 3. Divergence (JSD) of random sets of documents by cluster size.

summed over all clusters  $i$  where  $n_i$  is the size of cluster  $i$ .

Other studies that have measured within-cluster textual coherence include Braam et al. (1991) and Jarneving (2007a), although both used different mathematical formulations of coherence.

### *Concentration of grant-to-article linkages*

One of the challenges of comparing text-based and citation-based cluster solutions of a particular corpus is to find a metric that is independent of both text and citation, and that can be considered unbiased. Although only citation-based work is reported in this article, our complete study included both types of approaches. Given that a textual coherence is likely to favor text-based solutions over citation-based solutions, we needed a second accuracy measure, and one that was less biased toward either text or citation. In informal conversations about this study it was suggested<sup>8</sup> to us that the grant acknowledgements mined from MEDLINE might be a suitable dataset from which to design such a metric. A grant-to-article linkage dataset from a previous study (Boyack, 2009a), consisting of a matched set of grant numbers and PMID, was available for such a purpose.

In order to measure concentration, one must limit the basis set to those elements that can actually show a concentrated solution. For example, grants that have only produced one article cannot differentiate between cluster solutions. Thus, we limited the grant-to-article linkage set to those grants that have produced a minimum of four articles. The resulting basis set thus consisted of 571,405 separate links between 262,959 unique articles and 43,442 NIH grants.

<sup>8</sup> Dr. Bob Schijvenaars of Collexis, Inc. made the suggestion.

The premise for using these grant-to-article linkages as a metric for measuring the accuracy of a cluster solution is the assumption that the papers acknowledging a single grant should be highly related, and should be concentrated in a cluster solution of the document space. Using this approach, a cluster solution giving a higher concentration of grants would be considered to be more accurate than one with a lower concentration value. In addition, since grants are not inherently tied to the clustering of scientific articles either by text or by citations, we consider a grant-based metric to be unbiased.

We considered several different measures based on grant-to-article linkages including a standard Herfindahl (or concentration) index and precision-recall curves. The Herfindahl index had the advantage that it could be calculated on a grant-by-grant basis and then averaged over grants, thus ensuring high specificity. Its disadvantage is that it would give a single number for comparison. By contrast, a precision-recall method gives curves that show a distribution of metric values. The disadvantage of this approach is loss of grant-to-grant specificity; articles in a cluster might refer to different grants rather than the same grant.

We settled on the precision-recall measure so that we could see the distribution of the results. In this formulation of precision-recall one orders all clusters in a solution by the fraction of the articles in the cluster that reference a particular set of grants, and then generates a traditional precision-recall curve. In this case, recall would be the cumulative fraction of the links present in the ordered clusters (SumLink/TotLink in Table 4), and precision would be the cumulative fraction of the articles in the set of clusters retrieved that referenced the set of grants (SumArtL/SumArt in Table 4). In this setting, precision can be equated with concentration. This measure should be particularly well suited to comparing maps whose stated purpose is portfolio analysis since the portfolios that will be analyzed are most often tied to funded grants.

**Table 4. Example of cumulative precision-recall calculation based on grant-to-article linkages. Assume that the total number of linkages (TotLink) available is 2000.**

Clust	Art	ArtL	Links	Frac	SumArt	SumArtL	SumLink	R	Pr
1	100	90	150	0.90	100	90	150	0.075	0.900
2	100	80	130	0.80	200	170	280	0.140	0.850
3	100	70	120	0.70	300	240	400	0.200	0.800

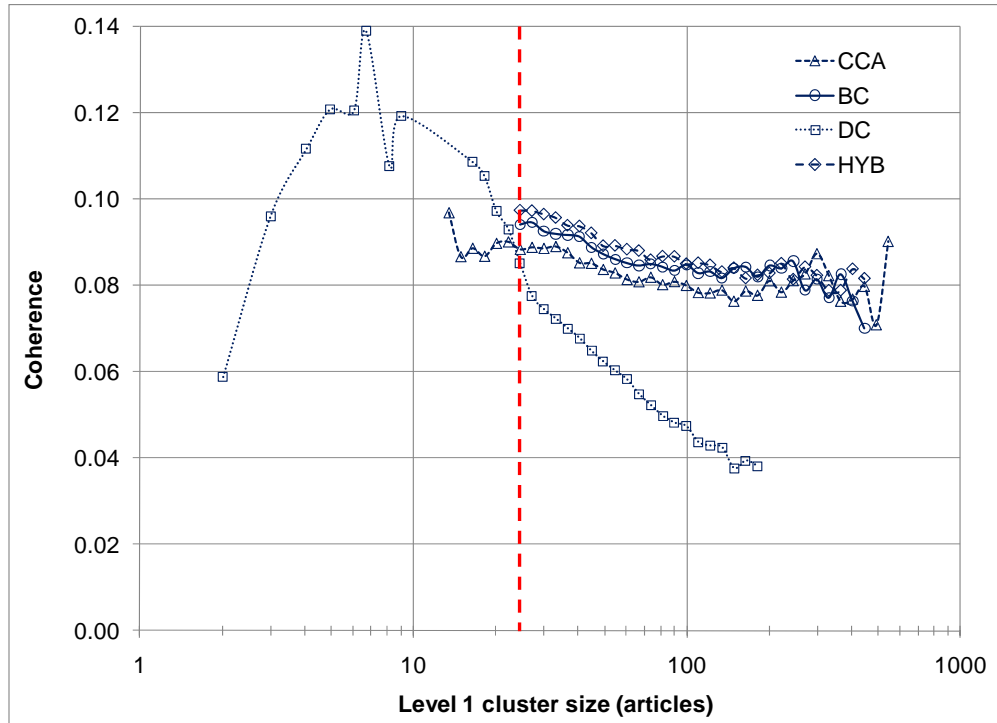
Art – number of articles in cluster, ArtL – number of articles in clusters linked to grants, Links – number of unique links to the ArtL (articles can be linked by more than one grant), Frac –  $\text{ArtL}/\text{Art}$ , Sum\* – cumulative sums, R – recall =  $\text{SumLink}/\text{TotLink}$ , Pr – precision =  $\text{SumArtL}/\text{SumArt}$ .

## Results

### *Coherence*

Since bibliographic coupling, direct citation, and hybrid solutions do not fractionalize articles, coherence for these approaches was calculated for using the full clustering results. However, co-citation analysis fractionalizes articles into multiple clusters. This feature cannot be fully captured by the coherence calculation. Thus, for purposes of comparison, coherence values for the co-citation solution were calculated using the assumption that each article can be uniquely assigned to its dominant (highest fractional value) cluster. Textual coherence is shown for the

four cluster solutions in Figure 4 as a function of cluster size. Each data point represents the average coherence for all clusters within the size bin.



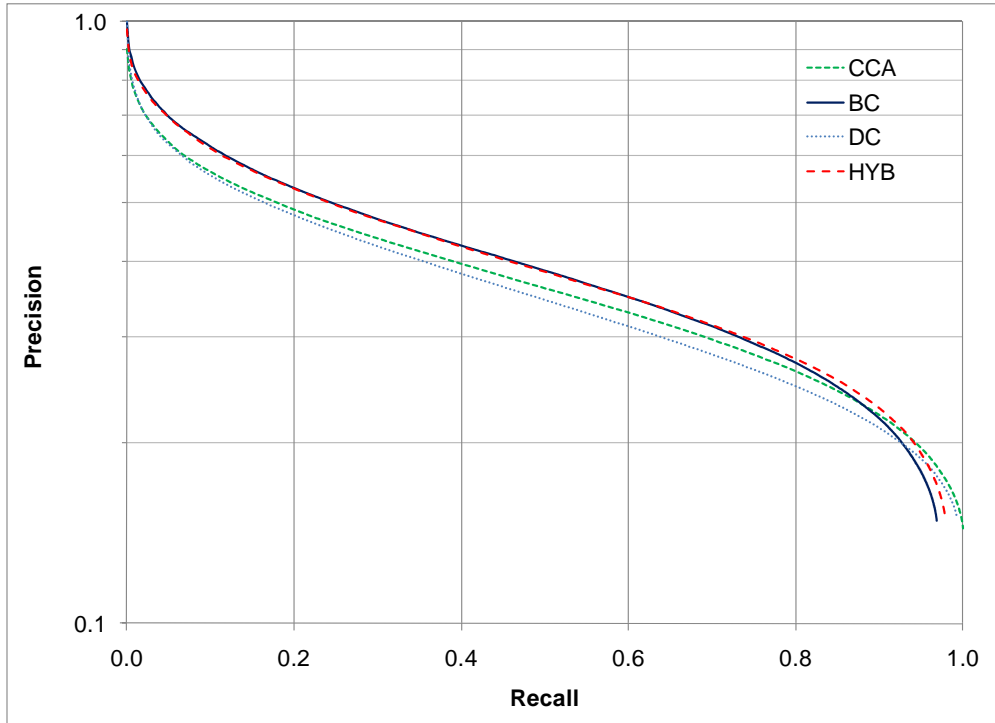
**Figure 4. Textual coherence values by cluster size for the four cluster solutions. Only those bins with 10 or more clusters are shown. The dashed line separates clusters with fewer than 25 articles from those with 25 or more articles.**

Three of the cluster solutions – co-citation analysis, bibliographic coupling, and the hybrid approach – show very similar results. For each curve, coherence decreases slightly with increasing cluster size. Although the differences between these three curves are small, they are measurable given that each curve represents over 30,000 clusters. Of the three approaches, the hybrid approach gives the highest textual coherence and thus the most accurate solution, followed closely by bibliographic coupling, and then by co-citation analysis. The direct citation solution also gives coherence values that decrease with increasing cluster size; however, the slope is greater and the coherence values much lower than for the other approaches. Although the direct citation approach does show higher coherence values for small clusters, these make up only a small part of the full cluster solution. Textual coherence values averaged over all articles were 0.0875, 0.0860, 0.0817, and 0.0614 for the hybrid, bibliographic coupling, co-citation analysis, and direct coupling solutions, respectively. We note that the hybrid solution might have been expected to give a higher textual coherence value simply because the hybrid is based partially on text components while the other three solutions are not.

### **Concentration**

Precision-recall curves were calculated for each cluster solution using the entire set of grant-to-article linkages mentioned above, and are shown in Figure 5. A higher precision value denotes a higher concentration of papers referencing the set of grants. The bibliographic coupling and

hybrid approaches have essentially identical curves up to a recall of 70% (0.7), after which the hybrid curve remains slightly higher than the bibliographic coupling curve. Both of these curves are higher than the co-citation analysis and direct citation curves up to a recall of over 90%. The co-citation analysis and direct citation curves are higher at the end because these two solutions cover a larger fraction of the 571,405 links than do the bibliographic coupling and hybrid solutions.



**Figure 5. Precision-recall curves for each cluster solution based on grant-to-article linkages. Results are limited to grants with four or more linked articles in the study corpus.**

Precision at 80% recall (Pr80) and the maximum value of F1 (a combined precision-recall statistic) are reported for each cluster solution in Table 5. The maximum F1 values for each solution occur at recall values between 0.56 and 0.60 for each of the solutions. Herfindahl index values (weighted by cluster size) for the solutions are also included in Table 5, and show that the solution order does not change using a different concentration measure, but is in fact accentuated; the bibliographic coupling and hybrid approaches are far superior to the other two using this measure.

It was also brought to our attention that different grant types have different properties that might behave differently in different cluster solutions, and thus could justify different analyses for different grant types. For example, single-component grants (e.g., NIH grant type R01) are typically focused on single topics. By contrast, multiple-component grants can include large center grants that focus on sets of related topics (e.g., NIH grant types M01, P01, P30) and equipment grants (NIH grant type P41) that may not be referenced in the same way that one would reference a topic-centered grant. Training grants (T32 type) are another grant type that may have very different publication patterns.

Precision-recall curves and statistics were calculated for R01, P01, and P41 types to compare to the overall results; Pr80 values for these are also in Table 5. The results for the R01 (single-component) grant type mirror the results for the full set, with the exception that the gap between the hybrid and bibliographic coupling approaches becomes larger (3.8% to 1.7%) at 80% recall. For P01 (multi-component) grants, the precision values for the four solutions are in the same order, and the gap between the hybrid and bibliographic coupling approaches is similar to that of the overall set (2.0% to 1.7%). These numbers suggest that a hybrid approach is more effective in a single-topic focus environment than in a multiple-topic environment.

**Table 5. Summary of concentration results for the different mapping approaches.**

Method	F1max (all)	Herfindahl	Pr80 (all)	Pr80 (R01)	Pr80 (P01)	Pr80 (P41)
CCA	0.4245	0.2378	0.2621	0.2147	0.0654	0.0423
BC	0.4410	0.2849	0.2706	0.2206	0.0692	0.0417
DC	0.4112	0.2037	0.2480	0.2004	0.0639	0.0460
HYB	0.4412	0.2893	0.2752	0.2290	0.0706	0.0420

The results for the P41 (equipment) grant type are very different; the most accurate approach for these equipment grants is direct citation. Pr80 values for the other three approaches are about 10% less than that for direct citation. This suggests that citing practices for articles acknowledging equipment grants are fundamentally different from citing practices for articles acknowledging topic-focused grants. We leave the reasons for these differences to future study.

## ***Discussion***

Defining which method for generating a map of the research front in the biomedical literature would give the most accurate results involves many dimensions. Of the possible dimensions involved in this identification we choose four to highlight here. Table 6 shows the values of four metrics for each of the cluster solutions: computational cost, coverage, textual coherence, and precision at 80% recall based on grant-to-article linkage information. These are important for different reasons:

- Computational cost – a low computational requirement is desirable for map generation, particularly since similarity value calculations typically scale with at least the square of the number of papers.
- Coverage – high coverage is necessary to allow for accurate portfolio analysis.
- Coherence – document clusters should be tightly focused in terms of their content
- Precision – document clusters should concentrate information from a grant-related standpoint to be more meaningful for portfolio analysis.

A comparison of the four similarity approaches shows that all three pure citation-based approaches have similar computational costs, while the hybrid approach has a higher computational cost. The hybrid approach took roughly 50% longer to calculate than did the bibliographic coupling approach due to the much larger number of co-occurrence values that needed to be calculated. Addition of citing:word pairs to the citing:cited pairs increased the total number of bibliographic coupling counts in the hybrid set by a factor of 3 over that for the pure bibliographic coupling approach. The co-citation approach had the highest coverage with 98.37% of the articles in the corpus being assigned to clusters. This was followed closely by the

bibliographic coupling and hybrid approaches with over 96.6% each. Given that coverage for all approaches was over 92%, and given that the differences in computation costs between the approaches are not prohibitive, we consider these to be the least important factors in differentiating between the mapping approaches.

**Table 6. Summary of results for the different mapping approaches.**

Method	Comp Cost	Coverage	Coh	Coh vs. BC	Pr80 (all)	Pr80 vs. BC
CCA	Low	<b>98.37%</b>	0.0817	-5.0%	0.2621	-3.1%
BC	Low	96.62%	0.0860	--	0.2706	--
DC	Low	92.68%	0.0614	-28.6%	0.2480	-8.4%
HYB	Medium	96.83%	<b>0.0875</b>	+1.7%	<b>0.2752</b>	+1.7%

Among the pure citation-based approaches, bibliographic coupling had the highest values using both the coherence and grant concentration metrics, and is thus considered the most accurate of the three approaches at representing the research front. Co-citation analysis was a close second, with coherence and concentration values only a few percent (5.0% and 3.1%, respectively) lower than those for bibliographic coupling. Direct citation is clearly the least accurate approach among those tested.

The hybrid approach, although computationally more expensive than bibliographic coupling, improved upon the bibliographic coupling solution in all respects – with higher coverage (by 0.2%) and higher coherence and concentration values (by 1.7%). Although these are modest gains, they establish the fact that the addition of textual information to citation-based approaches can increase their accuracy at the scale of millions of articles. We fully expect that experimentation with different hybrid formulations would lead to further gains in accuracy. The work by Liu et al. (2010) with a large number of hybrid formulations suggests this as well.

## Conclusions

In this study we sought to answer the question as to which citation-based mapping approach would generate the most accurate cluster solution (or science map) of the research front in the biomedical literature. To do this, we identified a large corpus (2.15 million articles) and generated cluster solutions using three standard citation-based approaches – co-citation analysis, bibliographic coupling, and direct citation – and one hybrid approach. A relatively complete corpus is necessary for portfolio analysis, the stated application of our work. Use of a smaller data set, or reduction of a large data set to ‘core’ documents (Jarneving, 2007a) would simply not be sufficient for this application.

Two different accuracy measures were used to compare the results from the four approaches. Of the pure citation-based approaches, bibliographic coupling gave the most accurate solution, followed closely by co-citation analysis. The bibliographic coupling-based hybrid had slightly higher accuracy than the pure bibliographic coupling approach.

This study also introduced a new approach to comparing the accuracy of cluster solutions. We used grant-to-article linkages from the grant acknowledgements in MEDLINE to calculate a precision-recall statistic showing how these linkages were more or less concentrated in different

cluster solutions. A cluster solution with a higher concentration of grant-to-article linkages is considered more accurate than a solution that provides a lower concentration. This method is based on the assumption that articles from a single grant should be concentrated together in a cluster solution. We fully recognize that this type of metric can only be used when grant-to-article linkage data are available, and that such data are relatively scarce except in the MEDLINE/NIH context.

Of the accuracy studies reviewed in the background section of this article, only two directly compared multiple citation-based mapping approaches in a quantitative fashion. The results of those two studies are compared with the results of this study in Table 7.

**Table 7. Comparison of results from accuracy studies of citation-based mapping approaches.**

Study	#Articles	Links used	Methods
Couto et al. (2006)	6,880	Internal, external for BC	Amsler > BC > CCC
Shibata et al. (2009)	23,459 (max)	Internal only	DC > BC > CCC
This study (2010)	2,081,022 (BC)	Internal, external for BC, CCA	BC > CCA > DC

CCC – co-citation clustering

Given that 1) the corpus used in this study is two orders of magnitude larger than that used in the larger of the two previous studies, 2) both internal and external linkages were used, and 3) full co-citation analysis was used instead of co-citation clustering, we consider the results of this study to be very robust. Among pure citation-based approaches, the research front is most accurately represented on a large scale by bibliographic coupling. Addition of textual information to the citation information in a bibliographic coupling approach increases the accuracy of the solution. These global findings do not exclude the possibility that for some local environments another approach may be more accurate. In fact, our results showing that direct citation was more accurate with regard to the literature associated with instrumentation grants is an example of such a local environment.

We note, once again, that the simple hybrid approach used here is only a first trial at a hybrid approach. The constraints used (a severe limiting of the word distribution) were done to keep processing time to a minimum and still explore the borders of the hybrid space. It is expected that additional experimentation would produce a hybrid approach with even higher accuracy. This study adds to the growing body of results at both the journal and document level showing that text-citation hybrid approaches have the potential to outperform approaches based solely on either text or citation. We will explore this idea further in future publications that report the results of the text-based approaches that were a part of our large NIH study.

Finally, we note that most of the data from this study, the list of PMID, titles and abstracts, similarity files, article-to-cluster assignments, and coherence results are available for download at <http://sci.slis.indiana.edu/sts/>. We invite others to use these data to make further comparisons; they should be very suitable for the development and testing of similarity approaches, clustering algorithms and accuracy measurement approaches.

## Acknowledgements

This work was supported by NIH SBIR Contract HHSN268200900053C. Russell Duhon, Indiana University extracted words from the titles and abstracts of documents in the corpus for use in the hybrid approach. We also acknowledge insightful comments and suggestions from the reviewers of the original manuscript.

## References

- Ahlgren, P., & Colliander, C. (2009a). Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, 3, 49-63.
- Ahlgren, P., & Colliander, C. (2009b). Textual content, cited references, similarity order, and clustering: An experimental study in the context of science mapping. *12th International Conference of the International Society for Scientometrics and Informetrics*, 862-873.
- Ahlgren, P., & Jarneving, B. (2008). Bibliographic coupling, common abstract stems and clustering: A comparison of two document-document similarity approaches in the context of science mapping. *Scientometrics*, 76(2), 273-290.
- Boyack, K. W. (2009a). Linking grants to articles: Characterization of NIH grant information indexed in Medline. *12th International Conference of the International Society for Scientometrics and Informetrics*, 730-741.
- Boyack, K. W. (2009b). Using detailed maps of science to identify potential collaborations. *Scientometrics*, 79(1), 27-44.
- Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351-374.
- Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis I. Structural aspects. *Journal of the American Society for Information Science*, 42(4), 233-251.
- Calado, P., Cristo, M., Goncalves, M. A., de Moura, E. S., Ribeiro-Neto, B., & Ziviani, N. (2006). Link-based similarity measures for the classification of web documents. *Journal of the American Society for Information Science and Technology*, 57(2), 208-221.
- Cao, M. D., & Gao, X. (2005). Combining contents and citations for scientific document classification. *AI 2005: Advances in artificial intelligence*, 143-152.
- Couto, T., Cristo, M., Goncalves, M. A., Calado, P., Ziviani, N., Moura, E., et al. (2006). A comparative study of citations and links in document classification. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 75-84.
- Glenisson, P., Glänzel, W., Janssens, F., & de Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing and Management*, 41, 1548-1572.
- Glenisson, P., Glänzel, W., & Persson, O. (2005). Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics*, 63(1), 163-180.
- Grosse, I., Bernaola-Galvan, P., Carpena, P., Roman-Roldan, R., Oliver, J., & Stanley, H. E. (2002). Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E*, 65, 041905.

- 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)*, 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.
- Jarneving, B. (2005). A comparison of two bibliometric methods for mapping of the research front. *Scientometrics*, 65(2), 245-263.
- Jarneving, B. (2007a). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, 1(4), 287-307.
- Jarneving, B. (2007b). Complete graphs and bibliographic coupling: A test of the applicability of bibliographic coupling for the identification of cognitive cores on the field level. *Journal of Informetrics*, 1, 338-356.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10-25.
- Klavans, R., & Boyack, K. W. (2006a). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2), 251-263.
- Klavans, R., & Boyack, K. W. (2006b). Quantitative evaluation of large maps of science. *Scientometrics*, 68(3), 475-499.
- Klavans, R., & Boyack, K. W. (2010). Toward an objective, reliable and accurate method for measuring research leadership. *Scientometrics*, 82(3), 539-553.
- Lin, J. (1991). Divergence measures based on Shannon entropy. *IEEE Transactions on Information Theory*, 37(1), 145-151.
- 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.
- Marshakova, I. V. (1973). A system of document connections based on references. *Scientific and Technical Information Serial of VINITI*, 6, 3-8.
- Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2008). *DrL: Distributed recursive (graph) layout*, SAND2008-2936J: Sandia National Laboratories.
- Sandström, U. (2009). *Bibliometric evaluation of research programs: A study of scientific quality. Report 6321*: Swedish Environmental Protection Agency.
- Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, 28, 758-775.
- Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology*, 60(3), 571-580.
- Sims, G. E., Jun, S.-R., Wu, G. A., & Kim, S.-H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences of the USA*, 106(8), 2677-2682.
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265-269.

- Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics*, 38(2), 275-293.
- Small, H., & Garfield, E. (1985). The geography of science: Disciplinary and national mappings. *Journal of Information Science*, 11, 147-159.