Identifying a Better Measure of Relatedness for Mapping Science

Richard Klavans
SciTech Strategies, Inc., 2405 White Horse Road, Berwyn, PA 19312. E-mail: rklavans@mapofscience.com

Kevin W. Boyack
Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185. E-mail: kboyack@sandia.gov

Measuring the relatedness between bibliometric units (journals, documents, authors, or words) is a central task in bibliometric analysis. Relatedness measures are used for many different tasks, among them the generating of maps, or visual pictures, showing the relationship between all items from these data. Despite the importance of these tasks, there has been little written on how to quantitatively evaluate the accuracy of relatedness measures or the resulting maps. The authors propose a new framework for assessing the performance of relatedness measures and visualization algorithms that contains four factors: accuracy, coverage, scalability, and robustness. This method was applied to 10 measures of journal–journal relatedness to determine the best measure. The 10 relatedness measures were then used as inputs to a visualization algorithm to create an additional 10 measures of journal–journal relatedness based on the distances between pairs of journals in two-dimensional space. This second step determines robustness (i.e., which measure remains best after dimension reduction). Results show that, for low coverage (under 50%), the Pearson correlation is the most accurate raw relatedness measure. However, the best overall measure, both at high coverage, and after dimension reduction, is the cosine index or a modified cosine index. Results also showed that the visualization algorithm increased local accuracy for most measures. Possible reasons for this counterintuitive finding are discussed.

Introduction

A variety of measures for journal, document, author, and word relatedness have been proposed and used in the literature (Jones & Furnas, 1987; McGill, Koll, & Noreault, 1979). Relatedness measures are necessary for a variety of reasons, from theoretical (e.g., gaining an understanding of the structure and dynamics of science) to practical (e.g., designing effective information retrieval and decision-support systems). Some researchers prefer focusing on intercitation (who cites whom) or cocitations (who is cited together in the same bibliography). Some are interested in the co-occurrence of words or authors. Some use simple measures such as raw frequency counts or normalized frequencies. Some prefer more computationally intensive methods such as Pearson correlations or chi-squares. Still others prefer to reduce the data into a two-dimensional (2-D) map, thereby creating an alternative measure of relatedness; the distance between tokens on a 2-D map is, in itself, a measure of relatedness.

Assessing the performance of these measures is critical for both theory development and practical application. As examples, insights into the structure or dynamics of science might be spurious if inaccurate measures are used. Information retrieval systems perform worse if less accurate measures of relatedness are used. We are particularly interested in use of these measures to assess and manage R&D. The use of inferior measures can result in a misallocation of R&D dollars, an action that can have serious economic, social, technological, and political consequences. It is the consequences of these decisions that drive our concern about the use of more accurate measures.

We focus on two questions that are basic to all science mapping efforts. First, how can we determine which relatedness (or similarity) measure is better from a pragmatic perspective? This is a timely question, given the recent criticism that the literature fails to emphasize the user's point of view (White, 2003). Second, can we determine how much performance is sacrificed when the data are reduced to two dimensions? This is also a timely question, given the recent emphasis on visualization (Börner, Chen, & Boyack, 2003; Chen, 2003) and the reasonable and common assumption that reduced accuracy goes hand in hand with reduced dimensionality.

We explore these questions in the context of journal–journal relatedness measures used for science mapping. We

Received August 2, 2004; revised January 11, 2005; accepted January 11, 2005

• Published online 11 November 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20274

begin with a brief background on commonly used similarity measures and derived similarity measures (using 2-D mapping algorithms). We then introduce a framework for assessing relatedness measures from a user's perspective. The framework consists of four criteria: accuracy, coverage, scalability, and robustness. We proceed to describe the data, relatedness measures, and additional data that we used to assess accuracy. We follow this with the results of the study along with their implications.

While we have developed this framework and the associated performance metrics as part of a larger project on developing new approaches to science mapping, we believe that they are very relevant to other applications in bibliometrics and data visualization where accuracy and validity matter.

Background

Relatedness Measures

Many different similarity measures are commonly used in bibliometrics. While we realize that word-based and citation-based measures are known to give different clustering results (Börner et al., 2003), we focus here on those measures that have had application in citation analysis because our study uses journal citation data. The two main groups of measures are intercitation measures, or those based on one journal citing another, and cocitation measures, which are based on the number of times two journals are listed together in a set of reference lists.

The simplest measure, raw frequency, is used for either intercitation counts or cocitation counts. Although raw frequency has been used for both journal citation (Boyack, Wylie, & Davidson, 2002) and journal cocitation analysis studies in the past (McCain, 1991), it is rarely used today. For intercitation studies, normalized frequencies such as the cosine, Jaccard, Dice, or Ochiai indexes (Bassecoulard & Zitt, 1999) are very simple to calculate, and give much better results than raw frequencies (Gmur, 2003). A new type of normalized frequency, specific to journals, has been proposed recently (Pudovkin & Fuseler, 1995; Pudovkin & Garfield, 2002). This new *relatedness factor* (RF), an intercitation measure, is unique in that it is designed to account for varying journal sizes, thus giving a more semantic or topic-oriented relatedness than other measures.

The Pearson correlation coefficient, known as *Pearson's r*, is a commonly used measure for journal intercitation (Leydesdorff, 2004a, 2004b), journal cocitation (Ding, Chowdhury, & Foo, 2000; McCain, 1992, 1998; Morris & McCain, 1998; Tsay, Xu, & Wu, 2003), document cocitation (Chen, Cribbin, Macredie, & Morar, 2002; Gmur, 2003; Small, 1999; Small, Sweeney, & Greenlee, 1985), and author cocitation studies (cf. White, 2003; White & McCain, 1998). Different authors treat the matrix diagonal differently—some leaving it as is and others treating it as missing data. Pearson's *r*, along with other statistical measures calculated from Pearson's *r*, such as chi-squares or *T* values, are also commonly used in genomics to calculate gene-pair relatedness (cf. Kim et al., 2001).

Other citation-based measures of relatedness include bibliographic coupling (Kessler, 1963) and combined linkage (Small, 1997). Bibliographic coupling suggests that two articles are related if they have common reference lists, while combined linkage combines direct citation counts with three types of indirect citations in a weighted average.

Visualization Methods

Lists of relatedness measurements are rarely analyzed directly, but are used as input to an algorithm that reduces the dimensionality of the data, and arranges the tokens on a 2-D plane. The distance between any two tokens on the 2-D plane is thus a secondary (or reduced) measure of relatedness. The most commonly used reduction algorithm is multidimensional scaling (MDS); however, its use has typically been limited to data sets of approximately tens or hundreds of items. Nonlinear MDS can deal with somewhat larger sets, around 10,000 nodes. Pathfinder network scaling (cf. Chen, Cribbin, Macredie, & Morar, 2002) is also used with smaller sets, allowing all of the links between items to be shown. Layout routines capable of handling more nodes include Pajek (Batagelj & Mrvar, 1998), which has recently been used to good effect by Leydesdorff (2004a; 2004b) on data sets with several thousand journals, the VxOrd graph layout routine (Davidson, Wylie, & Boyack, 2001), which has been used on a variety of data sets ranging into the tens of thousands of nodes (Boyack et al., 2002; Kim et al., 2001), and self-organizing maps (Kohonen, 1995), which can scale, with various processing tricks, to millions of nodes (Kohonen et al., 2000).

Factor analysis is another method for generating measures of relatedness. It is often used to show factor memberships on maps created using either MDS (McCain, 1998) or pathfinder network scaling (Chen et al., 2002). However, projections of two or three factors can be directly plotted and used to show relationships between objects. For instance, Leydesdorff (2004b) directly plotted factor values (based on citation counts) to distinguish between pairs of his 18 factors describing the *Social Science Citation Index* (*SSCI*) journal set. Factor analysis is best used when the number of descriptors is far less than the number of tokens, as in a recent study where it was used to classify a document set of 89,000 articles (tokens) and 887 common words (descriptors) in the field of genomics (Filliatreau et al., 2003). Factor analysis is not recommended for reduction of a square matrix; thus, it was not used in this study.

Validation of Relatedness Measures

Validation of relatedness measures has received little attention over the years. Most of these efforts have been to compare 2-D maps obtained from MDS with some sort of expert perceptions of the subject field. McCain

DOI: 10.1002/asi

(1986) compared the intellectual structures of two fields (macroeconomics and Drosophila genetics) from author cocitation analysis with the structures obtained from card-sorting surveys of authors in the fields. Similar studies using various expert elicitation methods include surveys (Perry & Rice, 1998) and interviews (Schwechheimer & Winterhager, 2001). In another case, the mental maps of 14 researchers were compared to bibliometric maps (Tijssen, 1993). In each of these cases, the citation-based maps were found to provide reasonable representations of the subject fields with respect to the expert opinions. In another study, Leydesdorff and Zaal (1988) compared dendrograms of 45 words from titles of biochemistry articles using four different co-word similarity metrics and found good agreement between the result from the different measures.

Only one study has compared citation-based relatedness measures. Gmur (2003) compared six different relatedness measures based on the cocitation counts of 194 highly cited documents in the field of organization science. The measures included raw frequency, three forms of normalized frequency, Pearson's *r*, and loadings from factor analysis. The bases for comparison were network-related metrics such as cluster numbers, sizes, densities, and differentiation. Results were strongly influenced by similarity type. For optimum definition of the different areas of research within a field, and their relationships, clustering based on Pearson's *r* or on the combination of two types of normalized frequency worked best.

We have found no previous work where the accuracy of different relatedness measures has been established quantitatively by comparison to a defensible standard. Our framework, methods, and results thus constitute the first such comprehensive study.

Proposed Framework

We propose a framework for choosing between different measures of relatedness that includes four criteria: accuracy, coverage, scalability, and robustness. Expected tradeoffs between these four criteria are discussed.

Accuracy

Accuracy refers to the ability of a relatedness measure to identify correctly whether tokens (e.g., journals, documents, authors, or words) are related. Accuracy in our context is analogous to the concept of precision in information retrieval. Assessments of accuracy can be conducted at two levels: local or global. Local accuracy refers to the tendency of the nearest tokens to be correctly placed or ranked. Ideally, local accuracy is measured from the perspective of each individual token. For authors, the question might be whether an author would agree with the ranking of the 10 most closely related authors. For journals, the question might be whether the closest journals were in the same discipline. For papers, the question might be whether the closest papers were on the same topic.

Global accuracy refers to the tendency for groups of tokens to be correctly placed or ranked, and requires that the tokens be clustered. A geographic analogy may help to explain the distinction between local and global accuracy. Local accuracy asks whether your immediate neighbors are correctly identified. Global accuracy assumes that towns exist (e.g., neighbors form clusters), and then focuses on whether the towns near you are correctly identified.

The assessment of accuracy requires some sort of independent data to use as a basis of comparison. One could use data from the perspective of each token (e.g., author rankings of which authors are most related, as in McCain's (1986) card-sorting study, or editor rankings of which journals are most related). One could also use data that represents the membership of each token (i.e., cliques of authors based on expert judgment, disciplinary groups of journals, or expert-based assignment of documents into research communities). To provide a basis for comparison, these data must be independent. For example, keywords should not be used if the tokens were words from the abstract of an article (one can expect that people use abstracts to assign keywords). However, keywords could be used to assess citation-based measures of document similarity (there is little evidence that citations are used by people assigning keywords).

In this article, we focus exclusively on *local accuracy*. The basis of comparison we use to establish accuracy in measures of journal–journal relatedness is the classification of journals from the Institute for Scientific Information (ISI; now Thomson ISI, Philadelphia, PA) journal categories. Any pair of journals is "related" if they belong to the same ISI category. It can be argued as to whether ISI provides the best available journal categorization. Yet, it has been constructed manually using both journal subject content and citation information (Morillo, Bordons, & Gomez, 2003; Pudovkin & Garfield, 2002), and thus represents a human judgment that can be considered as a high-quality standard of comparison. Independence from citation-based maps can also be argued, given that ISI is known to look at citation information as a part of their process for assigning journals to categories. However, given that the main purpose of this evaluation is to compare metrics, rather than to establish absolute accuracy, the ISI categories remain a suitable basis of comparison.

Coverage

Coverage helps to assess the impact of thresholds on accuracy. In this analysis, thresholds are used to identify all relationships that are at or above a certain level of accuracy. Very high thresholds of relatedness will tend to identify the relationship between a few tokens, lower thresholds will include more tokens, but the level of accuracy will likely be lower.

Coverage is here defined as the percentage of unique tokens that are identified for a specific threshold of relatedness. Thus, coverage in our context is analogous to the concept of recall in information retrieval. For example, a Pearson's *r* of 0.9 might only result in 500 of 7000 tokens

DOI: 10.1002/asi

being mentioned. A lower threshold of 0.6 might result in 5000 of 7000 tokens being mentioned.

Coverage is a valuable metric if one wants to compare the performance of different measures of relatedness. One might have a situation where one measure is more accurate for lower levels of coverage, and another measure is more accurate at higher levels of coverage.

There is a limit to coverage when citation-based measures are used. For example, at best, citation-based measures can only cover the full set of citing journals within a given data set. However, this is not the case for the cocitation-based measures. Cocitation-based measures can cover all of the cited journals (or conference proceedings) that are referenced within a given data set. It is known that there are many important journals or conference proceedings in the reference lists of papers that are not in the citing journal list (Tijssen & van Leeuwen, 1995). Cocitation measures can extend maps of science to include these journals where citation-based measures cannot.

Scalability

Scalability refers to the ability of a measure (or a derived measure from a visualization program) to be applied to extremely large databases. Some of the measures cannot be calculated for extremely large databases within reasonable timeframes. For example, applying Pearson correlations to journal data requires approximately n^2 calculations for n journals, and is extremely time consuming even with current computing capabilities. This is not a problem when one is dealing with smaller databases (less than 1000 tokens), but becomes intractable when one is dealing with very large databases (over 1 million tokens) because the response time is now measured in days. Very slow response times may be acceptable in academia, but many users require much faster response times.

Scalability is also an issue with visualization programs. Multidimensional scaling, the most popular approach, requires n^2 calculations. Alternatives, such as self-organizing map (SOM) and force-directed layout, use a variety of strategies to reduce the number of calculations to $n \log(n)$. These visualization programs run much faster, especially on extremely large databases (Börner et al., 2003).

Robustness

Robustness refers to the ability of a measure to remain accurate when subjected to visualization algorithms. Visualization algorithms reduce the dimensionality of the data, and it is reasonable to assume that the reduction in dimensionality will affect the accuracy of the measure. While the visualizations allow a user to gain insights into the underlying structure of the data, these insights should be qualified by an assessment of the concurrent loss of accuracy.

Tradeoffs

The relationships between scalability, coverage, accuracy, and robustness are important to consider. One expectation is that greater coverage will result in lower accuracy. For example, a relatedness threshold of 0.9 will probably identify journal pairs that are more accurate than a relatedness threshold of 0.6. Journal pairs with a threshold of 0.6 or more can be broken down into two groups: journal pairs with a threshold of 0.6 to 0.9, and journal pairs with a threshold of 0.9 to 1.0. It is reasonable to assume that the accuracy of the first group will be less than the accuracy of the second group.

Another expectation is that the measures that utilize more data and more calculations will be more accurate but less scalable. For example, we expect (a priori) the Pearson correlation to be most accurate (it uses almost the entire full matrix). However, the Pearson is not scalable at the level of 1 million tokens. Measures that are based on only a small segment of the full data matrix (such as frequencies or normalized frequencies) are probably less accurate (they use less information) but are more scalable.

A third expectation is that accuracy will drop when a measure is subjected to dimension-reduction techniques because the underlying data is inherently multidimensional. Dimensionality is reduced when specific measures of relatedness are applied. Dimensionality is further reduced when these measures of relatedness are used as inputs to visualization software. Each drop in dimensionality may correspond to a reduction in accuracy, and should be taken into account when interpreting the visual pictures.

The last tradeoff refers to the choice of intercitation versus cocitation measures. On the one hand, intercitation-based measures should be more accurate because the data are more current (current year to past years rather than past-year pairs). On the other hand, cocitation measures can cover far more sources. In this study, we limited our analysis to 7121 journals that are covered by ISI for the year 2000 (Thomson ISI, 2001a, 2001b). However, there are many non-ISI journals mentioned in the references of these articles (Leydesdorff, 2002), such as proceedings or regional and national journals. A cocitation measure has the potential of including thousands of additional journals into a map of science.

Data

The data used to calculate relatedness measures for this study were based on intercitation and cocitation frequencies obtained from the ISI annual file for the year 2000. *Science Citation Index Expanded* (*SCIE*; Thomson ISI, 2001a) and *Social Science Citation Index* (*SSCI*; Thomson ISI, 2001b) data files were merged, resulting in 1.058 million records from 7349 separate journals. Of the 7349 journals, we limited our analysis to the 7121 journals that appeared as both citing and cited journals. There were a total of 16.24 million references between pairs of the 7121 journals. Approximately 30% of all references could not be assigned to these 7121 journals. The resulting journal–journal citation frequency matrix was extremely sparse (98.6% of the matrix has zeros). While there was a great deal more cocitation frequency information, the journal–journal cocitation

DOI: 10.1002/asi

frequency matrix was also sparse (93.6% of the matrix has zeros).

We note that most previous studies of the relationship between journals have used data from the *Journal Citation Reports* (*JCR*) published by ISI. The *JCR* was not used here because, while it can be used for intercitation frequencies, it does not contain journal cocitation frequencies.

Additional data are required to measure accuracy. As mentioned previously, we used the ISI journal category assignments as the basis for comparison. For the combined *SCIE* and *SSCI*, there were a total of 205 unique categories. Including multiple assignments, the 7121 journals were assigned to a total of 11,308 categories, or an average of 1.59 categories per journal. There were 4019 journals that had a single category assignment, 2225 journals had two category assignments, and the remaining 877 journals had three or more assignments. For any journal pair, relatedness was considered to be (0, 1) binary: 1 if the two journals were assigned to a common category, and 0 if not. The ISI category assignments provide a matrix of comparable size to the calculated relatedness matrices (7121 × 7121), and that is similarly sparse (98.1% of the matrix has zeros).

Measures

We applied our framework and method to 10 different measures of journal–journal relatedness, six based on journal intercitation frequencies, and four based on cocitation frequencies. Given that most researchers do not analyze their relatedness measures directly, but use dimension reduction, we used these 10 measures as inputs to the VxOrd ordination algorithm (Davidson et al., 2001), effectively creating an additional 10 measures of journal–journal relatedness based on the distances between pairs of journals in 2-D space. We call these *re-estimated measures*. This second step allows us to determine which measure remains best after dimension reduction.

The VxOrd algorithm was chosen over MDS and other algorithms as the dimension-reduction routine for several reasons, some biased, and some practical. First, the algorithm was developed at Sandia National Laboratories (Albuquerque, NM), and we have had much experience with it. It has generated very useful data layouts (from an analyst's perspective) for a variety of (mostly unpublished) studies using different data sources. To us, a useful layout is one that has practical (but not perfect) fidelity at both the local and global scales; the local structure within clusters should make sense, and the relative placement of clusters should also make sense. On a more subjective basis, VxOrd is computationally efficient, using a density grid to model repulsive forces, with run times of order O(n). It has been used to generate graph layouts from data in excess of one million nodes and 8 million edges on a high-end PC, and thus is scalable to the graph sizes needed for the more granular models of science that we will generate in the future. We also note that the accuracy of other algorithms such as MDS has not been established for bibliometrics studies including thousands of nodes, and would welcome the appearance of such a study in the future.

The 10 relatedness measures used in this study are given below, along with their equations. The six intercitation measures are raw frequency, Cosine, Jaccard, Pearson's r, the recently introduced average *relatedness factor* of Pudovkin and Garfield (2002), and a new normalized frequency measure that we introduce here, K50.

IC-Raw $\quad RAW_{i,j} = RAW_{j,i} = C_{i,j} + C_{j,i},$

IC-Cosine $\quad \mathrm{COS}_{i,j} = \mathrm{COS}_{j,i} = \dfrac{RAW_{i,j}}{\sqrt{S_i\, S_j}},$

where $S_i = \sum_{j=1}^{n} RAW_{i,j,}$

IC-Jaccard $\quad JAC_{i,j} = JAC_{j,i} = \dfrac{RAW_{i,j}}{S_i + S_j - RAW_{i,j}},$

IC-Pearson

$$r_{i,j} = \frac{\sum_{k=1}^{n} (RAW_{i,k} - \overline{RAW_i})(RAW_{j,k} - \overline{RAW_j})}{\sqrt{\sum_{k=1}^{n} (RAW_{i,k} - \overline{RAW_i})^2 \sum_{k=1}^{n} (RAW_{j,k} - \overline{RAW_j})^2}},$$

where $\overline{RAW_i} = \frac{1}{n}\sum_{k=1}^{n} RAW_{i,k}, \quad k \neq i,$

IC-RFavg $\quad RFA_{i,j} = RFA_{j,i} = (RF_{i,j} + RF_{j,i})/2,$

where $RF_{i,j} = 10^6 * C_{i,j}/N_j\, S_i.$

IC-K50

$$\mathrm{K50}_{i,j} = \mathrm{K50}_{j,i} = \max\left[\frac{(RAW_{i,j} - E_{i,j})}{\sqrt{S_i\, S_j}}, \frac{(RAW_{i,j} - E_{j,i})}{\sqrt{S_i\, S_j}}\right],$$

where the expected value of the cosine $E_{i,j} = \frac{S_i S_j}{(SS - S_i)}$, and $SS = \sum_{i=1}^{n} S_i$.

Note that the new measure, K50, is simply the cosine index minus an expected cosine value. $E_{i,j}$ is an expected value of $RAW_{i,j}$, and varies with S_j, thus K50 is asymmetric and $E_{ij} \neq E_{ji}$. In each of the equations $C_{i,j}$ is the number of times journal i (fileyear 2000) cites journal j (all years), and N_i is the number of papers published in journal i in current year (in this case the 2000 fileyear). For all six intercitation similarity measures, we limited the set to those journal pairs for which $RAW_{i,j} > 0$. This is obvious for those measures with C_i or $RAW_{i,j}$ in their numerator, in that the calculated similarity will be zero for $RAW_{i,j} = 0$. However, this is not the case for the Pearson's r or K50, which often have non-zero results when $RAW_{i,j} = 0$. Note also that for our calculation of the Pearson correlations, we treat the diagonal as missing, a policy that is followed by most authors. A visual

DOI: 10.1002/asi

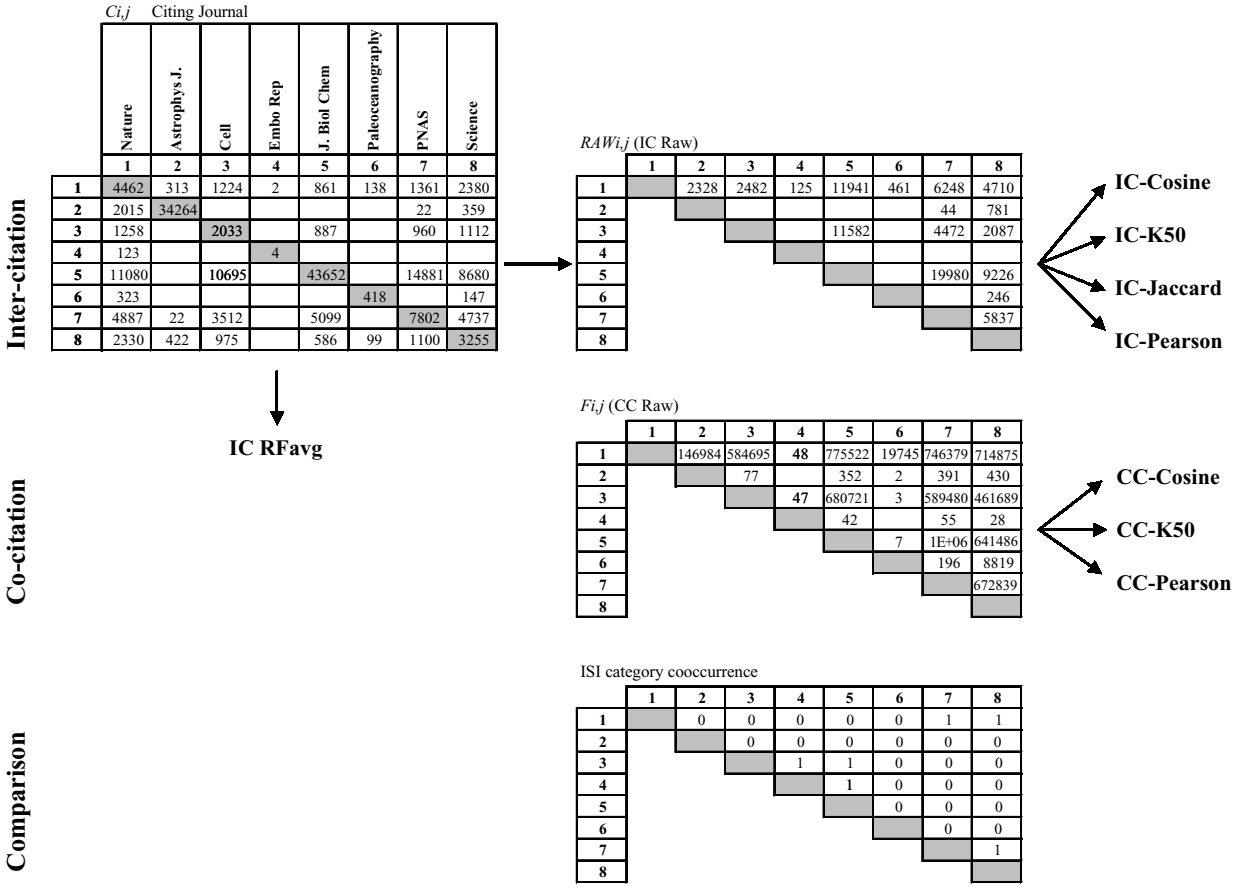

FIG. 1. Example illustrations of the intercitation $C_{i,j}$, $RAW_{i,j} = C_{i,j} + C_{i,j}$, cocitation ($F_{i,j}$) and ISI category co-occurrence matrices used in this study. Values are given for 8 of the 7121 journals from the ISI fileyear 2000 data: 1—*Nature*, 2—*Astrophysics Journal*, 3—*Cell*, 4—*Embo Reports*, 5—*Journal of Biological Chemistry*, 6—*Paleoceanography*, 7—*Proceedings of the National Academy of Sciences* (*PNAS*) *of the USA*, and 8—*Science*. Half-matrices are shown for the *RAW*, *F*, and ISI matrices since they are symmetric.

example of the $C_{i,j}$ and $RAW_{i,j}$ matrices is shown in Figure 1 for *Nature* and seven other journals.

The four cocitation measures are raw frequency, cosine, Pearson's *r*, and the cocitation version of the K50 measure.

CC-Raw $\quad F_{i,j}$,

CC-Cosine $\quad \mathrm{COS}_{i,j} = \mathrm{COS}_{j,i} = \dfrac{(F_{i,j})}{\sqrt{S_i S_j}}$,

where $S_i = \sum_{j=1}^{n} F_{i,j}$,

CC-Pearson $\quad r_{i,j} = \dfrac{\sum_{k=1}^{n} (F_{i,k} - \overline{F_i})(F_{j,k} - \overline{F_j})}{\sqrt{\sum_{k=1}^{n} (F_{i,k} - \overline{F_i})^2 \sum_{k=1}^{n} (F_{j,k} - \overline{F_j})^2}}$

where $\overline{F_i} = \frac{1}{n} \sum_{k=1}^{n} F_{i,k}, \quad k \neq i$,

CC-K50

$$\mathrm{K50}_{i,j} = \mathrm{K50}_{j,i} = \max \left[\frac{(F_{i,j} - E_{i,j})}{\sqrt{S_i S_j}}, \frac{(F_{j,i} - E_{j,i})}{\sqrt{S_i S_j}} \right],$$

where the expected value of the cosine $E_{i,j} = \frac{S_i S_j}{(SS - S_i)}$, and $SS = \sum_{i=1}^{n} S_i$.

In all four cocitation measures, $F_{i,j}$ is the frequency of co-occurrences of journal *i* and journal *j* in reference documents (from the combined reference lists of the fileyear 2000 data), and *n* is the number of journals. For the four cocitation measures, we limited the calculation to those journal pairs for which $F_{i,j} > 0$. A visual example of the $F_{i,j}$ (CC-Raw) matrix is given in Figure 1 for *Nature* and seven other journals, along with the ISI category assignment co-occurrence matrix used as the basis of comparison in this study.

Table 1 contains calculated values for all 10 relatedness measures for the *Nature-n* journal pairs from Figure 1, and shows some of the effects of different similarity measures. For instance, for small journals, the K50 values are nearly equal to the cosine values (see e.g., Paleoceanography), and thus small journals move up in the rankings. Conversely, the *Proceedings of the National Academy of Sciences of the United States* (*PNAS*) and *Science*, two well-known large multidisciplinary journals that are often associated in the same phrase with *Nature* are ranked in *Nature*'s top 4 for the IC-cosine, but they drop to being ranked 30 and 23,

DOI: 10.1002/asi

TABLE 1. Values of the 10 relatedness measures (and absolute rankings in parentheses) for the journal *Nature* paired with seven other journals (see Figure 1). Journals are sorted across the top by decreasing IC-Cosine. Values of N_i and S_i (intercitation row sum) for *Nature* are 3062 and 282,663, respectively.

	5—*J Biol Chem*	7—*PNAS*	8—*Science*	4—*Embo reports*	6—*Paleoceanography*	3—*Cell*	2—*APJ*
N_i	5592 (8)	2670 (25)	2595 (27)	92 (2666)	50 (4446)	347 (525)	2259 (34)
S_i (*IC*)	557773	361830	241764	282	4567	137980	162228
IC-Raw	11941 (1)	6248 (2)	4710 (3)	125 (498)	461 (111)	2482 (8)	2328 (12)
IC-Cosine	0.03007 (1)	0.01954 (2)	0.01802 (4)	0.01400 (12)	0.01283 (18)	0.01257 (19)	0.01087 (30)
IC-K50	0.01528 (1)	0.00762 (30)	0.00829 (23)	0.01367 (2)	0.01151 (5)	0.00525 (80)	0.00293 (246)
IC-Jaccard	0.01441 (1)	0.00979 (2)	0.00906 (3)	0.00044 (484)	0.00161 (99)	0.00594 (12)	0.00526 (14)
IC-RFavg	3.516 (481)	3.107 (561)	3.196 (543)	71.262 (1)	16.431 (10)	7.728 (104)	2.273 (768)
IC-Pearson	0.79700 (41)	0.92989 (2)	0.97618 (1)	0.12489 (841)	0.16199 (711)	0.89257 (3)	0.07349 (1104)
CC-Raw	775522 (1)	746379 (2)	714875 (3)	48 (3818)	19745 (121)	584695 (4)	146984 (14)
CC-Cosine	0.04724 (1)	0.04547 (4)	0.04708 (3)	0.00051 (1962)	0.01333 (28)	0.04714 (2)	0.01671 (19)
CC-K50	0.02486 (3)	0.02309 (4)	0.02643 (2)	0.00038 (525)	0.01136 (13)	0.03038 (1)	0.00490 (48)
CC-Pearson	0.90951 (19)	0.96030 (2)	0.99160 (1)	0.83943 (80)	0.26694 (1294)	0.95280 (3)	0.06810 (3723)

Note. *J Biol Chem,* Journal of Biological Chemistry; *PNAS*, *Proceedings of the National Academy of Sciences*; *APJ*, Astrophysics Journal.

respectively, by the IC-K50. The IC-RFavg tends to act in a different manner than all of the other measures, accentuating the (semantic) relationship between small and large journals, which was its intended effect (Pudovkin & Garfield, 2002).

As mentioned above, for each of the 10 relatedness measures, a dimension reduction was done using VxOrd. The process for calculating "re-estimated measures" is as follows. First, 2-D coordinates were calculated for each of the 7121 journals using VxOrd (cf. Figure 2). Next, the distances between each pair of journals (on the 2-D plane) were calculated for the entire set and used as the re-estimated measures of relatedness.

It is important to note that the full matrices were not used in the VxOrd step. We discovered during the validation phase that pictures that are more accurate could be generated if we used only the largest 15 similarities per journal. Thus, we culled the similarity files to include only the top 15 similarity pairs per journal, and these were used as input to VxOrd. Although this does exclude information from the journal network graph, using only the top *n* similarities can be justified by anecdote. An author, when deciding where to publish a particular paper, rarely considers more than just a few journals as an appropriate place to publish the work. With regard to that work, all other journals are irrelevant. Likewise, most journal publishers consider only a few other journals as close competitors, and worry very little about those outside that list. Thus, we feel very comfortable using only the dominant 15 links per journal in creating our maps of science. Indeed, a smaller number may be optimum, but we did not investigate this with parametric studies.

Analytical Results

Accuracy. The first factor in our framework for comparing different relatedness measures is accuracy. To provide a common basis for comparing relatedness measures with different distributional characteristics, we process the data in the following ways. First, ranked relatedness is used rather than absolute similarity values or distances (cf. Table 1). For each relatedness measure, the journal pair with the highest similarity value is assigned a rank of "1," the journal pair with the next highest similarity value receives a rank of "2," and so forth. At this level, we do not compare intercitation measures with cocitation measures because the total number of rankings is different. Using our calculation criteria, a total of 351,983 and 3,458,489 similarity values were calculated for the intercitation and cocitation measures, respectively.

Second, accuracy values were assigned to each of the ranked journal pairs for each similarity measure using (0, 1) binary relatedness from the ISI category assignments, as mentioned above. We plot cumulative accuracy because of the tendency to use thresholds in subsequent analyses. Cumulative accuracy tells us the average accuracy for all of the journal–journal pairs that meet or exceed a threshold.

DOI: 10.1002/asi

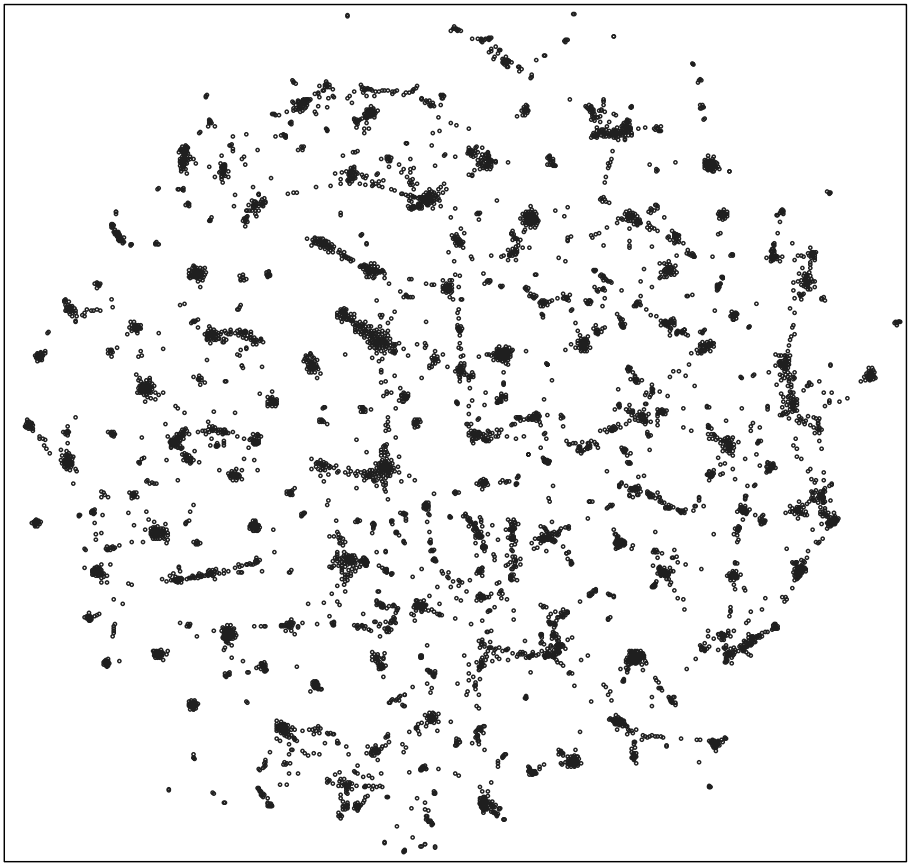

FIG. 2. VxOrd solution for 7,121 journals using the top 15 similarity values per journal and the intercitation cosine (IC-Cosine) measure.

Figure 3 illustrates the relationship between cumulative accuracy and ranked relatedness. For the intercitation measures (Figure 3a), there is the expected relationship between accuracy and ranked relatedness, with accuracy starting high and decreasing with increasing rank. The IC-Pearson measure is the most accurate for higher absolute levels of relatedness (up to a rank of ~85,000). As ranked relatedness increases, the curves for all but the IC-Raw measure converge. IC-Cosine, IC-K50, and IC-Jaccard measures generate nearly identical results over the entire relatedness range up to a rank of ~125,000. Raw citation frequencies provided the worst results over the entire range.

The cocitation measures (Figure 3b) have behavior similar to that of the intercitation measures, namely a brief period of volatility with high accuracy followed by a decrease in accuracy with increasing rank. The CC-Pearson measure is the best of the four up to a rank of ~350,000, and then drops below the CC-Cosine and CC-K50. The CC-K50 is slightly more accurate than the CC-Cosine, and the raw frequency measure, CC-Raw, gives the worst results by far.

Coverage. Plots of the relationship between coverage and ranked relatedness are shown in Figure 4, where coverage is defined as the number of unique journals represented at or above a specific rank. For example, for the IC-RFavg measure, a total of 3484 unique journals are named in the first 5000 ranked journal pairs. Figure 4a shows that for the intercitation measures, the IC-Cosine and IC-K50 measures cover more journals than the other measures over the entire range of rank relatedness. The IC-Jaccard and IC-RFavg measures have the next highest coverage, followed by the IC-Pearson. The IC-Raw covers the fewest journals over most of the range.

Figure 4b shows coverage results for the cocitation measures. The same pattern emerges. The CC-Cosine and CC-K50 have the highest coverage, followed by the CC-Pearson. Once again, raw frequency gives the worst results.

Accuracy and coverage. All measures of relatedness can be compared directly if one focuses on the tradeoff between cumulative accuracy and coverage (see Figure 5). "Accuracy versus coverage" in our context is analogous to the concept of "precision versus recall" in information retrieval. The most accurate raw measure is different at different levels of coverage. The IC-Pearson measure is more accurate for up to a coverage of 0.58, while the IC-Cosine and IC-K50 are more accurate for coverage past 0.58. The two raw frequency-based measures, IC-Raw and CC-Raw, are the two least accurate measures, peaking at 0.61 and 0.44, respectively, and have thus not been shown in Figure 5. Four remaining measures (IC-Jaccard, IC-RFavg, CC-Cosine,

DOI: 10.1002/asi

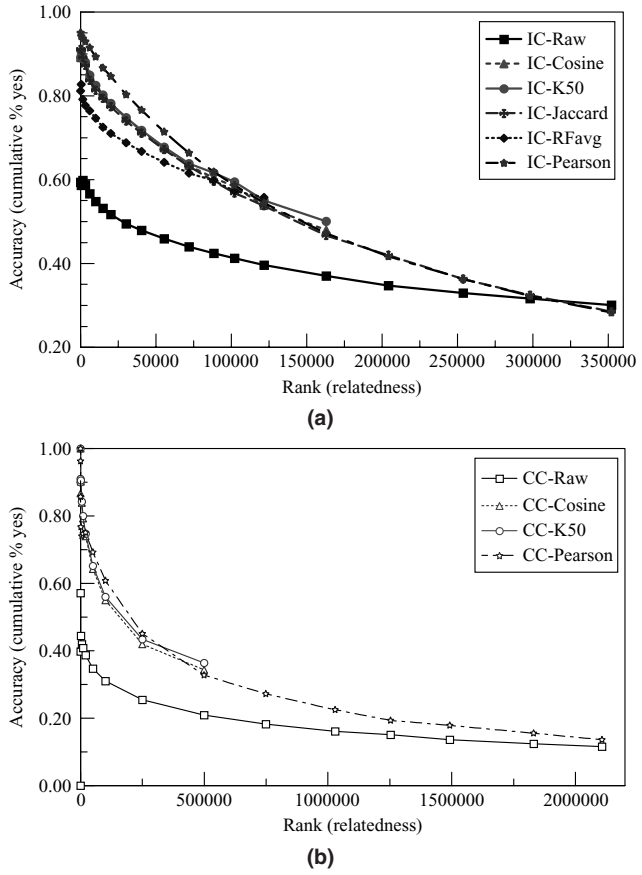

FIG. 3. Accuracy versus ranked relatedness for the (a) six intercitation measures, and (b) four cocitation measures.

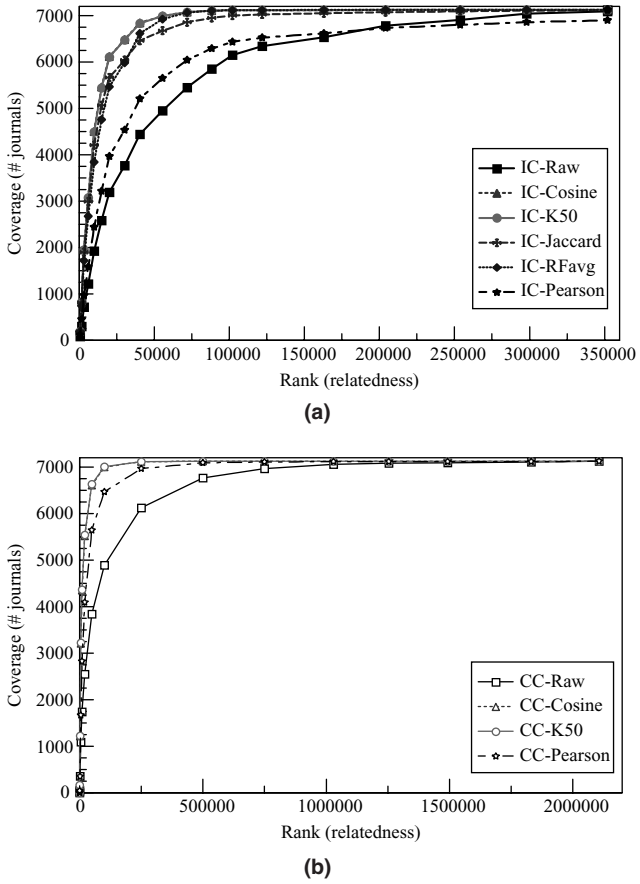

FIG. 4. Coverage versus ranked relatedness for the (a) six intercitation measures, and (b) four cocitation measures.

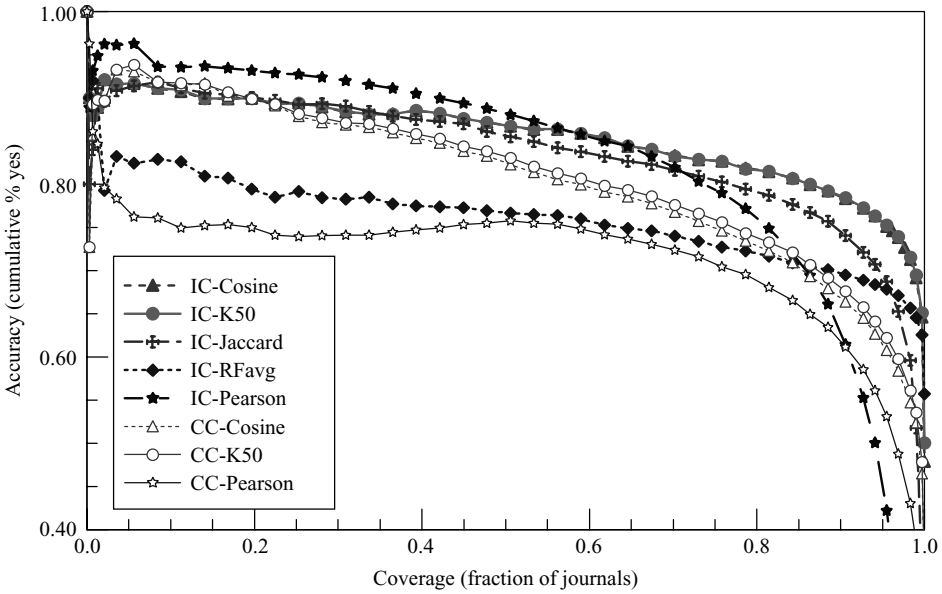

FIG. 5. Accuracy versus coverage curves for 8 of the 10 original relatedness measures. The IC-Raw and CC-Raw measures are not included here due to their low accuracy values.

and CC-K50) have comparable levels of performance that are less accurate than the best measures at high levels of coverage. Note that, excepting the raw frequency measures, both of which do poorly, the intercitation measures are more accurate than the cocitation measures.

Robustness. Accuracy and coverage were also calculated for the 10 different re-estimated (using the VxOrd ordination routine) relatedness measures. The intermediate step of converting 2-D distances between journal pairs into rank relatedness was done for these measures, but the plots are not

DOI: 10.1002/asi

shown here. For each relatedness measure, the journal pair with the shortest distance was assigned a rank of "1," the journal pair with the next shortest distance received a rank of "2," and so forth.

Figure 6 shows the accuracy and coverage tradeoff curves for eight of the re-estimated measures, and reveals several interesting things. First, the IC-Cosine, IC-K50, and IC-Jaccard measures all have roughly comparable accuracy over the entire range of coverage. The IC-K50 measure is slightly more accurate than the others from 20–50% coverage, while the IC-Cosine is the most accurate from 50–90% coverage. The IC-Pearson measure remains below these three over the entire coverage range. The IC-RFavg measure is the most consistent measure, maintaining roughly 85% accuracy over nearly its entire coverage range, and is the most accurate measure from 96–99% coverage (see inset in Figure 6). The IC-K50 measure is the most accurate above 99% coverage.

Second, the intercitation measures are more accurate than the cocitation measures in all cases. Third, the Pearson measures are less accurate than the cosine measures for both the intercitation and cocitation data. Also, note that the re-estimated K50 measures are essentially identical to the cosine measures for both the intercitation and cocitation data. Any differences at a particular coverage value are small enough to justify using the cosine value, which requires less calculation. It appears that, although the K50, by virtue of subtracting out the expected values, gives different individual similarity values and rankings, the aggregate effect on overall accuracy is minimal.

The most striking result comes from a comparison of the results of Figures 5 and 6, namely that *the overall accuracy for all re-estimated measures is higher than for the raw measures* over nearly the entire coverage range. This is an extremely counterintuitive finding, given the prevailing and common belief that information is lost when dimensionality is reduced. The marginal improvements in accuracy from the re-estimated measures are shown in Figure 7. Accuracy was reduced slightly for the IC-Cosine, CC-Pearson measure below 45% coverage, and for the IC-RFavg, CC-Pearson, and both K50 measures below 5% coverage. Accuracy was increased by the VxOrd procedure in all other cases. Notably, the visualization algorithm increased the accuracies of the IC-Cosine, IC-Jaccard, and IC-RFavg measures over the entire coverage range. We do not claim that all data sources or all dimension reduction techniques will show a similar improvement in accuracy with dimension reduction, but rather that it did for this combination. We do encourage further investigation into the quantitative effects of dimension reduction, particularly at the point of impact to the analyst.

A summary of the results of our investigation over the factors comprising our framework for comparing relatedness measures is shown in Table 2. Highlighted cells in the table show the measures with the best performance at different coverage levels. As mentioned above, the re-estimated measures provide better performance in nearly all cases, and thus will be used in making judgments between measures. We will also exclude any further discussion of the two raw frequency measures due to their overall poor performance.

Three of the intercitation measures (IC-Cosine, IC-K50, and IC-Jaccard) perform similarly, all with high-accuracy values at the both the 50% and 95% coverage levels. Given that the three are separated by only 1% accuracy at the 95% coverage level, it is our feeling that one would be justified in using any of the three if considering this alone. However, we see no reason to use the least accurate of the three, and thus would recommend usage of either the IC-Cosine or IC-K50 measures.

All of the intercitation measures are limited to use within the citing journal set. If coverage outside the citing journal

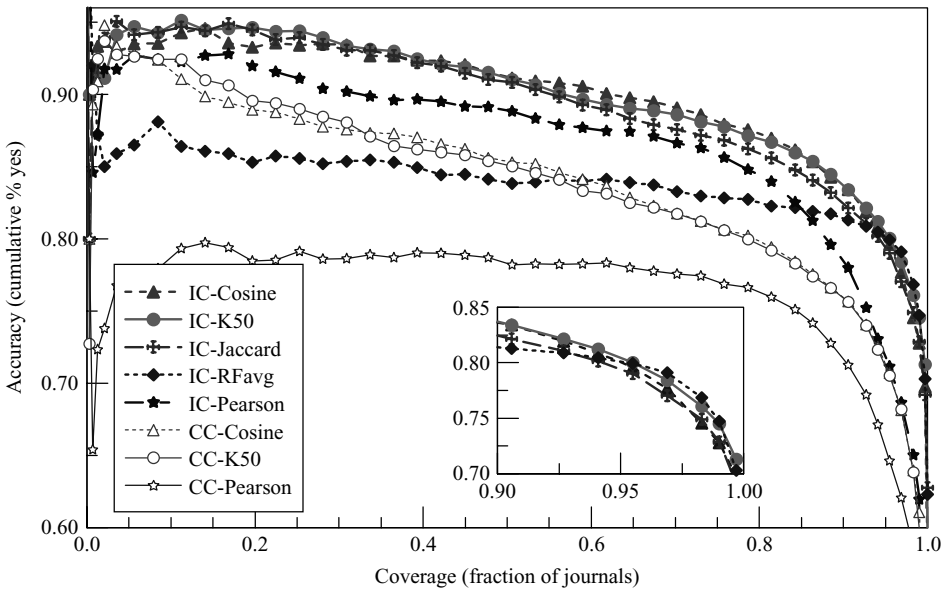

FIG. 6. Accuracy versus coverage curves for the re-estimated relatedness measures.

DOI: 10.1002/asi

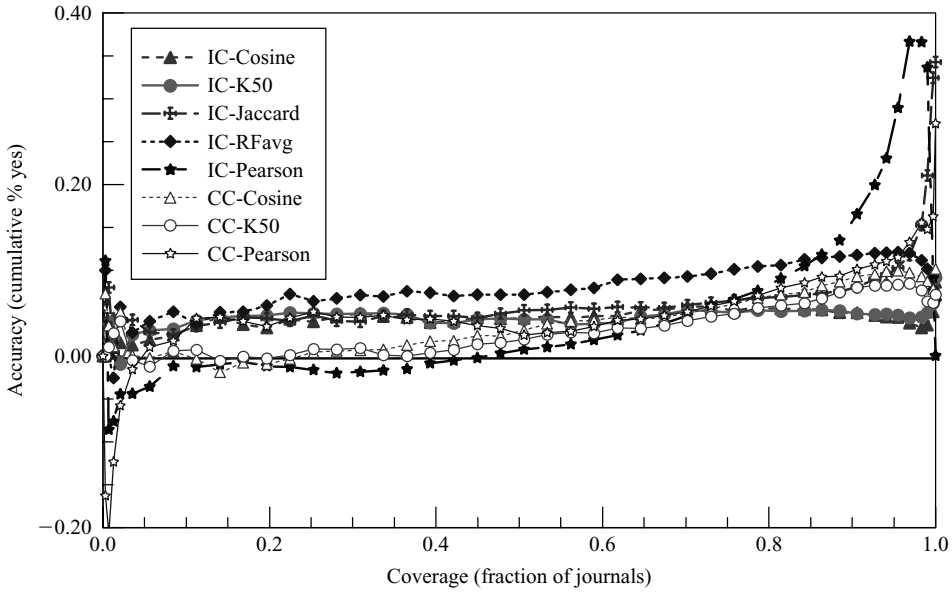

FIG. 7. Marginal improvement in accuracy when measures are re-estimated using VxOrd.

set is desired, cocitation measures can be used. Of these, the new measure introduced in this paper, CC-K50, is slightly better than the Cosine at high-coverage levels. Both the CC-Cosine and CC-K50 are clearly better than the Pearson correlation, both in terms of accuracy, and in that they do not require n^2 calculations, and thus scale to much larger sets than the Pearson.

Discussion

There were two results that were a surprise. First, we expected the Pearson correlation to provide the best results. The reason for this expectation is that the Pearson correlation uses more information in its construction (nearly the entire intercitation or cocitation matrix) than do the other measures. Pearson correlations allow for the influence of other parties. On the other hand, the other measures only use a small amount of the data in the matrix, and tend to limit their focus to the relationship between the two journals in question.

This is less of a surprise if one focuses on the conditions of low coverage where Pearson has an advantage. This is precisely the situation where Pearson correlations are often used in bibliometrics. For example, they have been used in author cocitation analyses to show the relatedness between elite or highly cited authors. These studies rarely cover less influential or new authors in a field, and thus cannot claim to have high coverage of a field.

The second surprise was the increase in performance from the visualization software. We expected the performance to deteriorate due to the simple rule of thumb that reducing data to two dimensions requires tradeoffs that would result in lower accuracy. Indeed, we have found only one other documented case where accuracy improved with decreased dimensionality. De Chazal and Celler (1998) used neural networks for electrocardiogram (ECG) diagnosis, and found that "single network classifier accuracy tended to improve as more principle components were removed." They found the opposite effect with a multiple network classifier structure.

TABLE 2. Performance of relatedness measures within the comparison framework.

Measure	Accuracy @ 50% coverage	Accuracy @ 50% coverage after VxOrd	Accuracy @ 95% coverage	Accuracy @ 95% coverage after VxOrd	Maximum coverage	Scalability
IC-Raw	52.4%	60.6%	36.9%	60.1%	Citing journal set	High
IC-Cosine	86.8%	91.3%	75.5%	80.2%	Citing journal set	High
IC-K50	86.8%	91.2%	75.5%	80.5%	Citing journal set	High
IC-Jaccard	85.6%	90.8%	69.7%	79.5%	Citing journal set	High
IC-RFavg	76.7%	83.9%	68.0%	80.2%	Citing journal set	High
IC-Pearson	88.3%	88.8%	44.7%	71.7%	Citing journal set	Low
CC-Raw	35.8%	22.9%	20.9%	25.6%	Cited journal set	High
CC-Cosine	82.5%	85.3%	61.6%	71.2%	Cited journal set	High
CC-K50	83.3%	85.1%	62.8%	71.4%	Cited journal set	High
CC-Pearson	75.8%	78.5%	54.5%	65.3%	Cited journal set	Low

DOI: 10.1002/asi

We do not know what it is about these journal citation data or the VxOrd algorithm that gives rise to the increase in performance seen in this study. However, we venture a guess. The improvement in performance may be explained by the peculiarities of the VxOrd force directed algorithm. VxOrd balances attractive forces between nodes (the similarity values) with those of a repulsive grid that tries to force all nodes apart. It also cuts edges once the similarity-to-distance ratio falls below a threshold, and in most cases cuts about 50% of the original edges, thus leaving edges only where particularly strong similarities exist among a set of nodes. These dominant similarities are likely to be very accurate on the whole, and when concentrated by pruning the less accurate edges, may increase the overall accuracy of the solution.

VxOrd also employs boundary jumping (Davidson et al., 2001), thus allowing nodes that are trapped in a high-energy position to jump to a lower energy, and thus more locally accurate, position. To picture this effect, imagine two people with their arms interlocked trying to get them apart. The elbow for one person is blocked from being close to their body by the elbow of the other person. If one person then slides an arm out of this position, both people can have their elbows close to their bodies, a lower energy solution. In VxOrd, boundary jumping is what allows the two elbows to disengage each other and find their lower energy positions.

Another possible explanation for the increase in accuracy with dimension reduction is that given the inherent structure of the relatedness matrices, the eigenvectors of the matrices may be more robust than the variation underneath.

Conclusions and Implications

We have provided a methodology for comparing relatedness measures on a quantitative basis. The methodology requires two sets of data, one that is used to generate the relatedness measures, and another, independent source to test the accuracy and coverage of the relatedness measures. Accuracy and coverage are graphed to identify which measures are superior under what conditions. The best measures are contingent on the coverage. For high coverage using both raw and re-estimated measures, the Cosine and K50 measures using intercitation data are uniformly good choices.

It is important to point out, however, that the cocitation measures (CC-Cosine and CC-K50) will be superior if one wants to extend this analysis to additional journals not covered by ISI. The *SCI*/*SSCI* only cover about 7000 journals, and these journals only account for roughly 75% of the cited papers in this database. There are far more sources of publications (i.e., proceedings, technical reports, or national journals) that are important to science and technology that are not covered by ISI. The cocitation model would be necessary if the initial domain is expanded to include these additional sources that are important to scientific publication.

It is also important to note the unexpected results of reducing dimensionality and increasing performance. While this is puzzling, the result has a practical consequence. The resulting 2-D maps are actually more accurate than the data used to generate the map. The particular algorithm used here, VxOrd, seems to provide the best of two worlds—easy interpretability (because the data can be displayed in two dimensions), and greater accuracy.

We have focused on local accuracy and coverage with respect to relatedness measures. In subsequent work we will expand our focus to global accuracy, distortion effects from highly connected tokens (e.g., multidisciplinary journals), and expansion beyond ISI coverage. The sum result of all of these studies should lead to more accurate and useful maps of science.

We also note that this study on accuracy could have been conducted in many different ways. For instance, journals could have been mapped using author/institution co-occurrences, or even using text analysis techniques (one or many) over title words from articles from different journals. Additional similarity measures could have been included (Pearson excluding zeros or including diagonals, for example). The issue of dimension reduction algorithms for such studies remains open as well. Multidimensional scaling remains the algorithm of choice for many bibliometricians. The framework introduced here could be easily used for these other studies, and we would welcome comparative and follow-up studies on these and related issues. Any such studies should use the 7000+ journals in the ISI databases to enable comparisons on a common basis.

Acknowledgments

We thank Katy Börner, Peter Lane, Loet Leydesdorff, and anonymous reviewers for constructive comments on the manuscript. This work was supported by the Sandia National Laboratories Laboratory-Directed Research and Development Program. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

References

Bassecoulard, E., & Zitt, M. (1999). Indicators in a research institute: A multi-level classification of journals. Scientometrics, 44, 323–345.

Batagelj, V., & Mrvar, A. (1998). Pajek—A program for large network analysis. Connections, 21(2), 47–57.

Börner, K., Chen, C., & Boyack, K.W. (2003). Visualizing knowledge domains. Annual Review of Information Science and Technology, 37, 179–255.

Boyack, K.W., Wylie, B.N., & Davidson, G.S. (2002). Domain visualization using VxInsight for science and technology management. Journal of the American Society for Information Science and Technology, 53(9), 764–774.

Chen, C. (2003). Mapping scientific frontiers: The quest for knowledge visualization. London: Springer-Verlag.

Chen, C., Cribbin, T., Macredie, R., & Morar, S. (2002). Visualizing and tracking the growth of competing paradigms: Two case studies. Journal of the American Society for Information Science and Technology, 53(8), 678–689.

Davidson, G.S., Wylie, B.N., & Boyack, K.W. (2001). Cluster stability and the use of noise in interpretation of clustering. Proceedings of IEEE Information Visualization (pp. 23–30). Piscataway, NJ: IEEE.

DOI: 10.1002/asi

de Chazal, P., & Celler, B.G. (1998). Selecting a neural network structure for EGG diagnosis. Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 20(3), 1422–1425.

Ding, Y., Chowdhury, G., & Foo, S. (2000). Journal as markers of intellectual space: Journal cocitation analysis of information retrieval area, 1987–1997. Scientometrics, 47(1), 55–73.

Filliatreau, G., Ramanana-Rahary, S., Blanchard, V., Teixeira, N., Kerbaol, M., & Bansard, J.-Y. (2003). Bibliometric analysis of research in genomics during the 1990s. Paris, France: Observatoire des Sciences et des Techniques.

Gmur, M. (2003). An analysis and the search for invisible colleges: A methodological evaluation. Scientometrics, 57(1), 27–57.

Jones, W.P., & Furnas, G.W. (1987). Pictures of relevance: A geometric analysis of similarity measures. Journal of the American Society for Information Science, 38(6), 420–442.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14(1), 10–25.

Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., et al. (2001). A gene expression map for *Caenorhabditis elegans*. Science, 293, 2087–2092.

Kohonen, T. (1995). Self-organizing maps. New York: Springer.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., et al. (2000). Self organization of a massive document collection. IEEE Transactions on Neural Networks, 11(3), 574–585.

Leydesdorff, L. (2002). Indicators of structural change in the dynamics of science: Entropy statistics of the SCI Journal Citation Reports. Scientometrics, 53(1), 131–159.

Leydesdorff, L. (2004a). Clusters and maps of science journals based on bi-connected graphs in the Journal Citation Reports. Journal of Documentation, 60(4), 371–427.

Leydesdorff, L. (2004b). Top-down decomposition of the Journal Citation Report of the Social Science Citation Index: Graph- and factor-analytical approaches. Scientometrics, 60(2), 159–180.

Leydesdorff, L., & Zaal, R. (1988). Co-words and citations: Relations between document sets and environments. In L. Egghe & R. Rousseau (Eds.), Informetrics 87/88 (pp. 105–119). Amsterdam: Elsevier.

McCain, K.W. (1986). Cocited author mapping as a valid representation of intellectual structure. Journal of the American Society for Information Science, 37(3), 111–122.

McCain, K.W. (1991). Mapping economics through the journal literature: An experiment in journal cocitation analysis. Journal of the American Society for Information Science, 42(4), 290–296.

McCain, K.W. (1992). Core journal networks and cocitation maps in the marine sciences: Tools for information management in interdisciplinary research. Proceedings of the ASIS Annual Meeting, 29, 3–7.

McCain, K.W. (1998). Neural networks research in context: A longitudinal journal cocitation analysis of an emerging interdisciplinary field. Scientometrics, 41(3), 389–410.

McGill, M., Koll, M., & Noreault, T. (1979). An evaluation of factors affecting document ranking by information retrieval systems. Syracuse, NY: School of Information Studies, Syracuse University.

Morillo, F., Bordons, M., & Gomez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. Journal of the American Society for Information Science and Technology, 54(13), 1237–1249.

Morris, T.A., & McCain, K.W. (1998). The structure of medical informatics journal literature. Journal of the American Medical Informatics Association, 5(5), 448–466.

Perry, C.A., & Rice, R.E. (1998). Scholarly communication in developmental dyslexia: Influence of network structure on change in a hybrid problem area. Journal of the American Society for Information Science, 49(2), 151–168.

Pudovkin, A.I., & Fuseler, E.A. (1995). Indices of journal citation relatedness and citation relationships among aquatic biology journals. Scientometrics, 32(3), 227–236.

Pudovkin, A.I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. Journal of the American Society for Information Science and Technology, 53(13), 1113–1119.

Schwechheimer, H., & Winterhager, M. (2001). Mapping interdisciplinary research fronts in neuroscience: A bibliometric view to retrograde amnesia. Scientometrics, 51(1), 311–318.

Small, H. (1997). Update on science mapping: Creating large document spaces. Scientometrics, 38(2), 275–293.

Small, H. (1999). Visualizing science by citation mapping. Journal of the American Society for Information Science, 50(9), 799–813.

Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the Science Citation Index using co-citations. II. Mapping science. Scientometrics, 8, 321–340.

Thompson ISI. (2001a). Science citation index expanded. Philadelphia: Author.

Thompson ISI. (2001b). Social science citation index. Philadelphia: Author.

Tijssen, R.J.W. (1993). A scientometric cognitive study of neural-network research: Expert mental maps versus bibliometric maps. Scientometrics, 28(1), 111–136.

Tijssen, R.J.W., & van Leeuwen, T.N. (1995). On generalising scientometric journal mapping beyond ISI's journal and citation databases. Scientometrics, 33(1), 93–116.

Tsay, M.-Y., Xu, H., & Wu, C.-W. (2003). Journal co-citation analysis of semiconductor literature. Scientometrics, 57(1), 7–25.

White, H.D. (2003). Author cocitation analysis and Pearson's *r*. Journal of the American Society for Information Science and Technology, 54(13), 1250–1259.

White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. Journal of the American Society for Information Science, 49(4), 327–356.

DOI: 10.1002/asi