Scientometrics (2015) 105:1665–1677
DOI 10.1007/s11192-015-1610-x

# Statistical tests for 'related records' search results

**Charles H. Smith[1] · Patrick Georges[2] · Ngoc Nguyen[3]**

Received: 4 February 2015 / Published online: 2 June 2015


**Abstract** Related records searching, now a common option within bibliographic databases, is applied to an individual result record as a secondary way of refining the retrieval set obtained from the primary subject search operation. In one approach, an individual result record is linked to other article records on the basis of the number of references cited they share in common, the theory being that two articles that cite many of the same sources are likely to be highly similar in subject content. Results of the secondary search are usually displayed in the order of each item's actual number of commonly-shared references. In the present paper we suggest an improved way of ranking the results, employing statistical significance tests. We suggest two approaches, one involving a statistical test previously unknown in bibliometric circles, the binomial index of dispersion, and the other employing the more familiar centralized cosine measure; these turn out to produce nearly identical results. An example demonstrating the application of these measures, and contrasting such with the use of raw totals, is provided. In the example the results rankings are found to be only modestly (positively) correlated, suggesting that much information is lost to the user when raw totals alone are made the basis for ordering results.

✉ Charles H. Smith
charles.smith@wku.edu

Patrick Georges
pgeorges@uottawa.ca

Ngoc Nguyen
ngoc.nguyen@wku.edu

[1] University Libraries, Western Kentucky University, 1906 College Heights Blvd., Bowling Green, KY 42101, USA

[2] Graduate School of Public and International Affairs, University of Ottawa, Social Sciences Building, Room 6011, 120 University, Ottawa, ON K1N 6N5, Canada

[3] Department of Mathematics, Western Kentucky University, 1906 College Heights Blvd., Bowling Green, KY 42101, USA

**Keywords** Citation analysis · Bibliographic coupling · Binomial index of dispersion · Centralized cosine measure · Related records · Statistical measures

## Introduction

Bibliometrics as a study has undergone a considerable evolution since its early days, progressing through efforts to elucidate general statistical properties of authorship and referral [e.g., Lotka's law (Lotka 1926) and Bradford's law (Bradford 1934)], to citation analysis per se, and, most recently, among other directions, to related records searching.[1] There are two main approaches to related records (*aka* 'similar records', etc.) searching; both begin with the user having already executed a search in some bibliographic database on a subject of interest to him or her. Once the primary list of results is retrieved, individual records from it are then examined by the user who, on finding one of particular interest, executes a second search command to retrieve a more focused set of records related to it specifically.

The first approach alluded to above is based primarily or completely on title words or added keywords: that is, the 'related records' can be expected to share some of these with the subject record initially obtained. A rather different approach to related records searching employs a form of citation analysis. In this form of bibliographic coupling (used, for example, in the databases *Web of Science*, *BIOSIS Previews*, and *Cambridge Scientific Abstracts*) an individual subject article record is linked to other article records on the basis of the number of references cited they share in common, the theory being that two articles that cite many of the same sources are likely to be highly similar in subject content. Typically, once the secondary search is executed, results are presented onscreen in an order dictated by the absolute number of shared references identified. Such secondary searches are usually executed through a command icon labelled "view related records", "find similar articles", or some such.[2]

Each approach has its strengths and weaknesses. The keyword approach, for example, while intuitively clear to the user, depends on the ability of authors or indexers to identify keywords that are the most useful, but this is often difficult to do, especially if there are concepts involved that are hierarchical (*e.g.*, organismal systematics, or conceptual relations such as 'natural selection' to 'evolution'). The citation analysis approach plumbs a deeper form of relationship, though it is often true that authors do not refer in their writings

[1] The literature on bibliometrics in general is enormous; for quick accountings of the subject see Donohue (1974), Wolfram (2003), De Bellis (2009), and Haustein (2012). The same is true even for citation analysis alone; consult Borgman (1990), Moed (2005), and De Bellis (2009) for reviews. Importantly, however, much of the attention given to citation analysis has been in: (1) the scientometric context of the sociology of science: *e.g.*, to identifying ways of establishing schools of scientific endeavor, roles of key figures, subject trends, etc., and (2) the evaluation of various kinds of database inconsistencies or omissions (especially as related to impact factors). Much less attention has been given to the investigation of user-focused needs, though this is beginning to change (see, for example, Zhao and Strotmann 2014).

[2] In *Web of Science*, the command icon "View Related Records" is positioned on the right hand side of each of the records obtained through the primary search; the pop-up descriptor linked to the command reads "View other records that share references with this one". There is very little literature on the citation analysis form of related records searching beyond notices in product reviews (e.g., Wiley 1998), probably because the retrieval algorithms involved feature simple match counting.

to other works they probably should have; additionally, the approach is further removed from the appreciation of the ordinary user.

Here, our focus is to be on the citation analysis-based approach to related records searching only. We believe we can offer a refinement of the method that can be of service to the user by presenting a better-ordered results list.

## The problem

The method of establishing the bibliographic coupling relationship solely on the absolute number of shared items has a significant weakness: it does not take into account how many total references cited each of the articles has. Consider the Venn diagram presented as Fig. 1.

In this representation, the references cited portion of two articles ('1' and '2') is conceptually diagrammed as a pair of circles, with the size of each circle corresponding to the overall number of references cited in each. Article '1' cites 50 references, while article '2' cites 100. They share in common ten references cited; in the diagram this is indicated as the intersection set '*a*'. This leaves '*b*' and '*c*', those references cited in each article which are not cited in the other.

The question that immediately arises is what it may mean if for any given pairing of articles, '*a*' remains the same value, but '*b*' and '*c*' are both much smaller or much larger—or, for that matter, if one is much smaller and the other much larger. Also, what of the situation where '*a*' is 5, and '*b*' and '*c*' are, respectively, 20 and 45?: in this instance, the *proportion* of intersection is the same, but can we argue that this means the two articles are likely to be equally similar to the first example?

Thus, the problem is that the full relationship between '1' and '2' (that is, *any* '1' and '2') cannot be judged simply on the basis of '*a*' alone: we need to take into account both similarities and differences to get the entire picture. And, we need something more sophisticated then a Venn diagram to indicate how any such pairing of items in the whole database relates to it as a whole.

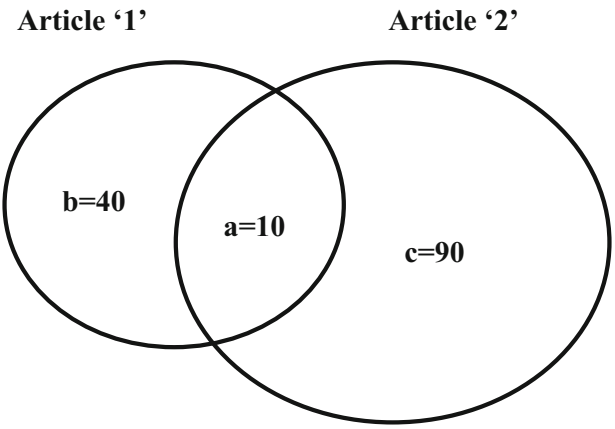


**Fig. 1** Common and distinct citations from articles '1' and '2' (subset *a* represents the 'intersection set' of commonly-held citations, whereas *b* and *c* represent distinct citations originating from each of the two articles)

This kind of problem is frequently encountered in biological systematics, where the objective may be to determine how similar two species populations are to one another, or two faunas (in a biogeographic sense) are to one another.[3] To these situations, statistics called 'measures of association' (*aka* 'similarity indices', and some other names) are often applied. In one of the most familiar of these (upwards of a hundred have been suggested over the years), the Jaccard index (Jaccard 1901), a statistic is generated from '*a*', '*b*', and '*c*' by calculating, for the numbers of forms in any pairing of regions, the value: $a/(a + b + c)$, which serves to proportionalize the numerical relationship between the numbers of shared and not shared forms. Oftentimes the matter of interest is not just one such comparison, but the relations among all pairings of five or ten geographical regions, and accordingly a matrix of values covering all combinations is produced.

In the bibliometric context, a migration toward the use of such a measure to relate the meaning of shared references would itself represent an advance, but we suggest a different approach, as the use of measures of association can itself be problematic. As conceived, they are almost all proportional measures, making it easy to effect comparisons among entirely different sets of data. But this approach comes at a cost. Potential information is in fact lost; the situation is similar to what happens when for a particular data set, descriptive coefficients of variation are related instead of independent means and variances. Each of the many measures of association that have been proposed, such as the Jaccard index above, have idiosyncrasies related to their particular formulations, which in effect add weightings of one kind of another to the relationships among '*a*', '*b*' and '*c*'. Thus, for the exact same raw data, a variety of different index scores will result, depending on the particular formulation involved; further, within a particular matrix of scores even the rank size order of same may vary depending on the formulation.

We have been able to identify two roughly equivalent ways of improving on this situation within the context of related records searches. These involve measures that are not merely proportional indices, and that can be applied to produce significance tests for the statistics generated by bibliographic coupling relations within a particular database. We will now describe these in turn.

## The binomial index of dispersion

### Formulation

Recently, two of us (Smith and Georges 2014) investigated the matter of record couplings in the context of a data set we developed that is closely analogous to the present related records question (see further discussion below). We came upon an index we felt we could use which is known as the binomial index of dispersion. In order to describe the index in a bibliographic context, we need to introduce some notation. Suppose a set $C$ of articles (potential citations) in the database so that the cardinality $(C) = |C| = n$, the number of articles in the database. For any pair of articles $(i, j)$ for $i, j \in C$ (among the $n \times n$ possible pairs), we are interested in capturing whether an article $k \in C$ had a reported influence (is a citation) on both $i$ and $j$, on $i$ but not $j$, on $j$ but not $i$, and on neither $i$ nor $j$. Running this across all articles $k$ for each pair $(i, j)$ we eventually obtain the set $I_i$ of all citations in

[3] Regarding biological measures of association, two well-known reviews are Cheetham and Hazel (1969) and Hayek (1994). As far as we can tell such measures have not been used in the past to contribute to a database user-oriented citation analysis mission.

**Table 1** 2 by 2 frequency table for citations using counts (see text for explanation)

| | Article $j$ | | | |
|---|---|---|---|---|
| Article $i$ | | Presence | Absence | Total |
| | Presence | $a$ | $b$ | $a + b$ |
| | Absence | $c$ | $d$ | $c + d$ |
| | Total | $a + c$ | $b + d$ | $n$ |

article $i$, and the set $I_j$ of all citations in $j$. Also, for any pair $(i, j)$, $I_i \cap I_j = CI_{i,j}$ is the set of articles $k$ that are cited in both $i$ and $j$; $I_i - I_i \cap I_j = I_{i,-j}$ is the set of articles $k$ that are cited in $i$ but not in $j$; $I_j - I_i \cap I_j = I_{j,-i}$ is the set of articles $k$ that are cited in $j$ but not in $i$ and $DI_{i,j} = I_{i,-j} \cup I_{j,-i}$ is the set of articles $k$ that are cited in either $i$ or $j$ but not both. This will eventually produce a count table (see Table 1) for the pair $(i, j)$ that sums the elements (the number of articles) in each of the four sets $CI_{i,j}$, $I_{i,-j}$, $I_{j,-i}$, and $C - CI_{i,j} - DI_{i,j}$, and from which a similarity index for the pair of articles $(i, j)$ can be computed on the basis of well-known formulas (e.g., the Jaccard index above, etc.). In what follows we discuss only the binomial index of dispersion, as this arguably provides the generally most useful kinds of results and is based on the $\chi^2$ statistic. This index can be computed for any pair $(i, j)$ as:

$$\mathrm{BID}_{i,j} = n(ad - bc)^2/[(a + b)(c + d)(a + c)(b + d)], \quad (1)$$

where $a$, $b$, $c$, $d$, and $n$ are the count/number of articles in each of the five sets $CI_{i,j}$, $I_{i,-j}$, $I_{j,-i}$, $C - CI_{i,j} - DI_{i,j}$ and $C$ (see again Table 1). Eventually, this method generates $n^2$ indices, one for each of the $n \times n$ pairings of articles $(i, j)$.[4]

The binomial index of dispersion is based on the $\chi^2$ statistic. Using the notation in Table 1, note that when two articles are independent (lack of association), the proportion or frequency of joint influences $(a/n)$ is equivalent to the product of the proportions $(a + b)/n$ and $(a + c)/n$ (that is, the proportion of articles in the database that are cited in $i$ and the proportion of articles cited in $j$). Therefore: $a/n = ((a + b)/n) \times ((a + c)/n)$ or equivalently $a = (a + b)(a + c)/n$. If the observed frequency is greater than the one expected under independence, then the two articles may be said to be positively associated. Thus, if articles $i$ and $j$ are associated, then: $a \neq (a + b)(a + c)/n$, and the difference could be written as:

$$D = a - (a + b)(a + c)/n = (a/n)(n - a - b - c) - bc/n = (ad - bc)/n. \quad (2)$$

This term $D$, or some variation of it, is found in the formula for calculating the usual Chi square statistic (Eq. 1) and all of its monotonically related statistics.

Intuitively, the binomial index gives, for any pair of articles, an index value that could be used as a simple check on to what degree two articles are associated. As noted above, if $D$ is not zero, the two articles of concern are associated. A larger value of $D$ would indicate a stronger association. In practice, the universe of all articles could be considered as the population, and thus the particular database of article records, as regard to a particular field of study, can be considered as a random sample from the population. As a result, the

[4] Strictly speaking the order of two articles in each pair $(i, j)$ is not important (i.e., the index is symmetrical), and we do not need to compare an article $i$ with itself. Hence there are only $n(n-1)/2$ relevant indices.

difference $D$ might be non-zero due to randomness or actual association between article records $i$ and $j$. Thus, if a pairing of article records results in a value of $D$ which is close to zero, the non-zero difference $D$ might just be due to randomness. We therefore need a way to assess whether $D$ is significantly different from zero. Fortunately, the Chi square statistic is equipped with the possibility of a statistical significance test. Under the condition that all expected frequencies in the presence/absence table (which is computed assuming independence of articles) are at least five and the sample size is sufficiently large, the distribution of the Chi square statistic is asymptotically Chi square distributed with one degree of freedom. The Chi square test of independence can then be used to assess whether there is a statistically significant association between two articles. A concrete example will be given later (as described in Table 2) when we discuss a specific database.

## Context

It is instructive to examine the circumstances under which the binomial index has been applied in the past. The index was introduced through the work of Potthoff and Whittinghill (1966), but has been applied somewhat infrequently over the years. Part of the reason for this was a change during that period in the goals of biosystematics work. Whereas prior to this time the assessment of relationships between species populations focused on simple presence–absence characteristics across an array of measured character states, a new approach dwelling on phylogenetic assumptions started to take over at this time. Thus, biosystematics turned more in the direction of establishing evolutionary order of derivation instead of mere morphological distinction. The index has instead found its main use in a different kind of setting. It was designed to test for homogeneity of proportions and has been applied in various contexts in the literature, but especially in epidemiology where it detects disease clustering by comparing proportions of cases among different areas. Examples of this can be found in Bassanezi et al. (2003), Wallet and Gotway (2004), and Spósito et al. (2007). In education, Rogosa et al. (1984) used the index to test for homogeneity of proportions of target teachers' behavior for Bernoulli-trial data, and thereby assessed the stability of teachers' behavior over time. Derivations and generalizations of the index have also been applied. For example, Duncan and Duncan (1955) used the square root of the index in computing and comparing measures of concentration of spatial variations in social phenomena, and Brandyberry et al. (1999) used the multinomial index of dispersion in a management science context to test for company size-related biases in different geographic locations. For examples of the use of the binomial index when two proportions are compared as a measure of association in biogeography utilizing presence/absence tables, see Cheetham and Hazel (1969) and Hayek (1994). Through presence/absence tables, frequency of joint presence or positive match, frequency of joint absences or negative match, and frequency of mismatches of states, habitats, species, factors, or variables are recorded. Based on this type of table, the significance of association between factors or variables can be established against the hypothesis of independence or random association.

To our knowledge, the binomial index has not been previously applied to any situation that could be interpreted as bibliometric in nature.

Thus, summarized, the binomial index accomplishes not only the result of taking into account two differing-sized data sets, but it establishes in any given pairing (i.e., as related to the universe of comparisons across the entire set of data sets) the degree of probability that '$a$' represents an unlikely outcome.

## The centralized cosine measure

A solution to the problem at hand can also be achieved through statistics more familiar to the bibliometrics community, involving the cosine similarity measure. The basic idea was mathematically formalized by Sen and Gan (1983); Glänzel and Czerwon (1996) later extended and applied the methodology. To find the relationship between two articles, each article $i$ is treated as a $n \times 1$ vector in the space of all $n$ articles in the database. If an article $k$ among the $n$ articles was cited on $i$, then the $k$th component of the vector corresponding to article $i$ obtains the value 1.0, otherwise it obtains a value of 0. Then, with respect to all articles in the database, each article is represented by a Boolean vector of 0's and 1's. The cosine similarity measure for a pair of articles $(i, j)$, say, $A_i$ and $A_j$, can then be computed by:

$$\mathrm{COS}_{i,j} = \frac{\sum_{k=1}^{n} A_{k,i} \times A_{k,j}}{\sqrt{\sum_{k=1}^{n} \left(A_{k,i}\right)^2} \sqrt{\sum_{k=1}^{n} \left(A_{k,j}\right)^2}}, \tag{3}$$

where subscript $k$ in $A_{k,i}$ indicates the $k$th component of vector $A_{k,i}$.

Thus, the cosine of the angle between the two vectors gives what is, in essence, a measure of similarity. The value of cosine similarity ranges between 0 and 1.0, where 0 indicates orthogonality or complete difference between two articles, and 1.0 indicates that two articles are exactly identical. In terms of association between two articles, values 0 and 1.0 both indicate dependence between two articles. Hence a value somewhere in the middle of the 0–1.0 range indicates degrees of independence of two articles. By computing the cosine similarity measure for any pair of articles in the database, we can then have some idea about the relationships among articles.

The cosine similarity measure is also known as Salton's measure, defined as the ratio of the number of joint references to the geometric mean of the number of references in the two articles (Glänzel and Czerwon 1996). Using this definition and our notations in the frequency table for citations using counts (Table 1), the formula for the cosine measure can be rewritten as:

$$\mathrm{COS}_{i,j} = \frac{a}{\sqrt{(a+b)(a+c)}}. \tag{4}$$

Comparing Eqs. (1) and (4), a clear difference between the binomial index and the cosine measure is recognized. While the binomial index takes into account the size of the database ($n$), the cosine measure does not.

The next step is to find a statistical test for the cosine index. Unfortunately, in our study all the vectors are Boolean vectors, so the null distribution of the cosine similarity under the assumption of independence between two articles is unknown and has a nonzero mean; in order to derive a statistical test for the cosine measure, a centralized cosine measure was proposed (Giller 2012). The centralized cosine measure is the cosine measure computed on the centralized vectors, with respect to the mean (average) vectors. Assuming that: $\overline{A_i} = \frac{1}{n}\sum_{k=1}^{n} A_{k,i}$ and $\overline{A_j} = \frac{1}{n}\sum_{k=1}^{n} A_{k,j}$, the centralized cosine measure CSC is:

$$\mathrm{CSC}_{i,j} = \frac{\sum_{k=1}^{n} \left(A_{k,i} - \overline{A_i}\right) \times \left(A_{k,j} - \overline{A_j}\right)}{\sqrt{\sum_{k=1}^{n} \left(A_{k,i} - \overline{A_i}\right)^2} \sqrt{\sum_{k=1}^{n} \left(A_{k,j} - \overline{A_j}\right)^2}}. \tag{5}$$

Values of the centralized cosine measure range from $-1.0$ to 1.0. A value of 1.0 indicates that two articles are identical. A value of $-1.0$ indicates that two articles are completely different. A value of 0 shows that two articles are independent (unassociated). Similar to the argument regarding the binomial index, the nonzero value of the centralized cosine measure might be due to randomness or actual association between articles. Thus a statistical significance test is required. Under the assumption that the size of the database $n$ is large enough, the distribution of the centralized cosine measure (under the assumption of independence) is approximately normal, with mean 0 and variance $1/n$. The centralized cosine measure can then be used to assess whether there is a statistically significant association between two articles.

Note that both binomial index and centralized cosine measure can be used to judge statistical significance of the association between two articles and both significant tests are asymptotic and require large database size.

## An example of application of these measures

Unfortunately we cannot deliver an example of the use of the index in a typical bibliographic context: the data we would need are of a proprietary nature, and unavailable to us. However, we can provide an example of its application in a directly analogous situation. This relates to the earlier-mentioned analysis of Smith and Georges (2014).

The object of that study was to investigate whether one might be able to assess degrees of similarity among classical music composers on the basis of the composers who influenced them. Data of the latter type were collected to form part of Charles Smith's reference website "The Classical Music Navigator" ('*CMN*'; Smith 2000). One can envision the comparisons made in the context of Fig. 1, where '1' and '2' now represent two composers in the database, '*a*' represents other composers in the database who influenced *both* '1' and '2', and '*b*' and '*c*' represent additional composers in the database who influenced '1' or '2', but not both at once. The analogy to the bibliographic coupling context is not perfect, but it is arguably strong enough to make some points about the inadvisability of using '*a*'-values alone as a measure of degree of association, and to suggest an interesting resemblance between similarity rankings obtained from the binomial index and the centralized cosine measure.[5]

From the Smith and Georges analysis we have pulled the lists of composers that the binomial index identified as being the 'most related' on this basis to the top four most frequently influencing composers in the *CMN* database, namely Stravinsky, Debussy, J. S. Bach, and Wagner. Table 2 details the situation for the Romantic era composer Richard Wagner; column 2 of Table 2 arranges the 26 ($=\hat{n}$) composers statistically most similar to him (these 26 are identified only by capital letters, their actual identities being unimportant here) in the rank and absolute order of the Chi square values that were obtained. Note that the table contains many ties, so the rank values given represent averages across what would otherwise be, without the ties, multiple values.

[5] A primary objective of the present work is to introduce database providers to the possibility of a new kind of tool, but ultimately it will be up to them to adapt the ideas to their own circumstances. The statistical approach itself may of course be applied to any setting—including humanities subjects—that meet the basic conditions of order within the data.

**Table 2** Comparisons of raw totals ('*a*') and statistical measures for Richard Wagner's "most similar" composers using the binomial index and cosine measures

| Influencing composer (coded by letter) | 1. Raw total rank (and actual total '*a*') | 2. Binomial statistic rank (and actual statistic) | 3. Centralized cosine measure rank (and actual statistic) | 4. Ordinary cosine measure rank (and actual statistic) |
|---|---|---|---|---|
| A | 4.5 (7) | 1 (170.4) | 1 (0.584) | 1 (0.592) |
| B | 4.5 (7) | 2 (147.6) | 2 (0.543) | 2 (0.553) |
| C | 14 (6) | 3 (145.7) | 3 (0.540) | 3 (0.548) |
| D | 4.5 (7) | 4 (129.9) | 4 (0.510) | 4 (0.522) |
| E | 14 (6) | 5 (106.7) | 5 (0.462) | 5 (0.474) |
| F | 24 (5) | 6 (99.5) | 6 (0.446) | 6 (0.456) |
| G | 39 (4) | 7 (96.8) | 7 (0.440) | 7.5 (0.447) |
| H | 14 (6) | 8 (93.7) | 8 (0.433) | 7.5 (0.447) |
| I | 4.5 (7) | 10 (86.4) | 10 (0.416) | 10 (0.434) |
| J | 4.5 (7) | 10 (86.4) | 10 (0.416) | 10 (0.434) |
| K | 4.5 (7) | 10 (86.4) | 10 (0.416) | 10 (0.434) |
| L | 24 (5) | 12 (84.1) | 12 (0.410) | 16 (0.423) |
| M | 14 (6) | 14.5 (83.3) | 14.5 (0.408) | 13.5 (0.424) |
| N | 14 (6) | 14.5 (83.3) | 14.5 (0.408) | 13.5 (0.424) |
| O | 14 (6) | 14.5 (83.3) | 14.5 (0.408) | 13.5 (0.424) |
| P | 14 (6) | 14.5 (83.3) | 14.5 (0.408) | 13.5 (0.424) |
| Q | 1 (8) | 17 (79.5) | 17 (0.399) | 17 (0.422) |
| R | 39 (4) | 20 (76.0) | 20 (0.390) | 22 (0.400) |
| S | 39 (4) | 20 (76.0) | 20 (0.390) | 22 (0.400) |
| T | 39 (4) | 20 (76.0) | 20 (0.390) | 22 (0.400) |
| U | 39 (4) | 20 (76.0) | 20 (0.390) | 22 (0.400) |
| V | 39 (4) | 20 (76.0) | 20 (0.390) | 22 (0.400) |
| W | 14 (6) | 23.5 (74.8) | 23.5 (0.387) | 18.5 (0.405) |
| X | 14 (6) | 23.5 (74.8) | 23.5 (0.387) | 18.5 (0.405) |
| Y | 24 (5) | 25.5 (72.5) | 25.5 (0.381) | 25.5 (0.395) |
| Z | 24 (5) | 25.5 (72.5) | 25.5 (0.381) | 25.5 (0.395) |

To this are added the parallel scores based on the centralized cosine measure [column 3, representing Eq. (5)] and the ordinary cosine (or Salton) measure [column 4, representing Eq. (3)]. As shown in column 2 of the table, the binomial (Chi square) statistic for Richard Wagner and composer *R* is 76.0. Using the Chi square distribution, the critical value at a 5 % significance level is 3.84. (for significance levels at 1 or 10 %, the critical values are 6.63 and 2.70, respectively) Because $76.0 > 3.84$, we can reject the null hypothesis of no association between Wagner and "*R*" in favor of the alternative that these two composers are statistically significantly associated. Equivalently for the cosine measures, given in column 3, the centralized cosine measure for Richard Wagner and composer *R* is 0.39, which is not close enough to 0 or 1 to provide an immediate judgment of significant association between two composers. The corresponding *z*-statistic for the centralized cosine value of 0.39 is 8.716, which is greater than the critical value of 1.96 at a 5 % significance level under the standard normal distribution. We can then reject the null

hypothesis of no association between Wagner and "*R*", which agrees with the conclusion using the binomial index.

One intriguing point is that the rankings produced by the binomial index and by the centralized cosine measure are exactly the same for the greater portion of significant values, with the statistical significance cut-off point for both measures at 5 % being at the same composer (not reported in Table 2 as this composer is at the 173rd position). Only at much lower critical values do we begin to observe some variation in rankings across both measures. Note also in Table 2 that the ordinary cosine measure does not generate the same rankings as the centralized cosine measure (or the binomial index).

This synonomy of rankings at first puzzled us, but it can be shown that there is a quadratic relationship between the BID and CSC measures. Note that in terms of our notation in the presence-absence table:

$$\sum_{k=1}^{n} A_{k,i} = a + b$$

$$\sum_{k=1}^{n} A_{k,j} = a + c$$

$$\sum_{k=1}^{n} A_{k,i}A_{k,j} = a$$

$$\overline{A_i} = \frac{a+b}{n}$$

$$\overline{A_j} = \frac{a+c}{n}.$$

After some algebraic simplification, we reach

$$\mathrm{CSC} = \frac{\sum_{k=1}^{n}(A_{k,i} - \overline{A_i}) \times (A_{k,j} - \overline{A_j})}{\sqrt{\sum_{k=1}^{n}(A_{k,i} - \overline{A_i})^2}\sqrt{\sum_{k=1}^{n}(A_{k,j} - \overline{A_j})^2}} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

From this, it can be easily seen that

$$\mathrm{BID} = n\mathrm{CSC}^2. \tag{6}$$

Thus the relationship between CSC and BID is not linear. However, when $n$ is large and CSC is close to 1, as in the case of two articles which are very similar, the relationship is close to a linear relationship.[6] This explains why the two measures are highly correlated for the top composers in our lists.

[6] Also, given Eq. (6) and because CSC can take negative values, BID is not a monotonic function of CSC. Hence, the order (or ranking) between CSC and BID is not preserved over the full set of CSC values. In our data set where negative values for CSC always remain very close to zero, then both BID and CSC will always generate the same ranking for relatively similar composers—that is, when CSC takes a positive value not too close to zero. Note that the ranking based on BID is equal to the ranking of the set of absolute values of CSC. The overall nature of this synonomy has led us to consider some related statistical and philosophical issues bearing on the measure and meaning of similarity and relatedness in data sets of the present type. We hope to explore this further in a future publication.

We have also placed in Table 2, as column 1, the values of the corresponding '*a*' totals, and their ranks. It can easily be seen that column 1 correlates rather imperfectly, and barely significantly, with the other three: the Pearson's *r* correlation statistic between the vectors of raw and statistical data in columns 1 and columns 2 and 3 is 0.422; the parallel value between column 1 and column 4 is 0.490. The matching *r* statistics for the ranks vectors are 0.470 and 0.591, respectively. For the remaining three composers, the parallel values are (in the same order): Stravinsky ($\hat{n} = 59$): 0.533/0.574/0.506/0.538; Debussy ($\hat{n} = 37$): 0.822/0.774/0.546/0.571; J. S. Bach ($\hat{n} = 31$): 0.604/0.603/0.435/0.513.

Some interesting discrepancies stand out in the instance of Wagner; note especially that the seventeenth ranked statistic in column 2 matches with the column 1 raw total—(i.e., '*a*'-) ranked first (composer 'Q'). Conversely, the composer ('G') ranked seventh in column 2 is only ranked thirty-ninth as a raw ('*a*') total. The reason for these discrepancies lies mainly in the varying lengths of the lists of composers identified as influences: in the case of composer 'G', only four influences on his work were identified—but all four of these turn out to be influences on Wagner as well, suggesting a close association. In the case of composer Q, there were more composers identified as influencing both him and Wagner—eight—than in any other such matchup, but overall 18 composers were identified as having been direct influences on him, so there were ten that were not matched with Wagner. Thus the basic laws of probability identify what might initially seem to be somewhat surprising results.

## Conclusion

It is our hope that database providers will consider the preceding discussion, and perhaps shift their 'related records' algorithms in a direction featuring use of the binomial index or cosine measure. We cannot claim at present that these latter necessarily provide a 'better' appraisal of relatedness—there seems to be no way of proving this, as 'relatedness' is at its core a philosophical concept, not a mathematical one—but it does seem possible to argue that they produce statistically superior representations of the data relationships involved. Note that Table 2 represents the results in terms of the statistics, but these could just as easily be expressed as level-of-significance values.

Meanwhile, it is useful to allude to some possible directions for further research. One question that can be posed is whether it is preferable to use, as a basis for the binomial calculations, all the records in a database (or random samples therefrom), or deliberately restricted portions of these (as for example in the *Web of Science*, in which are identified several dozen general subject categories with which its millions of item records are associated). There is also the matter of how small a representative sample would be satisfactory as a basis for standardizing the intra-data relations, and how often this might have to be calculated to keep the statistic reliable. Further, and as Chi square tests are normally not applied to datasets involving very small numbers of cases, some minimum cutoff value for number of references cited in an article might need to be set for it to be considered in relation to others.

Additionally, there is at least one users-oriented question to deal with. This is, simply, will they be willing to depend on a statistic they may not understand to assess relatedness, in preference to the less complex '*a*'-based tally? At the very least, a brief explanation would have to be placed directly at hand; possibly, the '*a*' value might also be listed in the search results.

Consider also that the more logical mathematical basis for a 'related records' function might be productively turned back on more general questions related to research, and the sociology of science. For example, studies looking into cross-discipline fertilization should profit from a more rigorous way of establishing related links.

Finally, it should be reiterated that significance measures for related records retrievals aside, even the switch to a measure-of-association index value to compare articles would itself likely represent an improvement over the '*a*'-based indicator.

## References


Bassanezi, R. B., Filho, A. B., Amorim, L., Gimenes-Fernandes, N., Gottwald, T. R., & Bové, J. M. (2003). Spatial and temporal analyses of citrus sudden death as a tool to generate hypotheses concerning its etiology. *Phytopathology, 93*(4), 502–512.

Borgman, C. L. (1990). *Scholarly communication and bibliometrics*. Newbury Park, CA: Sage Publications.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering: An Illustrated Weekly Journal (London), 137*, 85–86.

Brandyberry, A., Rai, A., & White, G. P. (1999). Intermediate performance impacts of advanced manufacturing technology systems: An empirical investigation. *Decision Sciences, 30*(4), 993–1020.

Cheetham, A. H., & Hazel, J. E. (1969). Binary (presence–absence) similarity coefficients. *Journal of Paleontology, 43*(5), 1130–1136.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the Science Citation Index to cybermetrics*. Lanham, MD: Scarecrow Press.

Donohue, J. C. (1974). *Understanding scientific literatures: A bibliometric approach*. Cambridge, MA: MIT Press.

Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review, 20*(2), 210–217.

Giller, G. L. (2012). The statistical properties of random bitstreams and the sampling distribution of cosine similarity. *Giller Investments Research Notes (20121024/1)*. http://dx.doi.org/10.2139/ssrn.2167044. Accessed 17 January 2015

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional, and institutional level. *Scientometrics, 37*(2), 195–221.

Haustein, S. (2012). *Multidimensional journal evaluation: Analyzing scientific periodicals beyond the impact factor*. Berlin: De Gruyter/Saur.

Hayek, L.-A. C. (1994). Analysis of amphibian biodiversity data. In W. R. Heyer, M. A. Donnelly, R. W. McDiarmid, L.-A. C. Hayek, & M. S. Foster (Eds.), *Measuring and monitoring biological diversity: Standard methods for amphibians* (pp. 207–269). Washington, DC: Smithsonian Books.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*(142), 547–579.

Lotka, A. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(12), 317–324.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Potthoff, R. F., & Whittinghill, M. (1966). Testing for homogeneity. I. The binomial and multinomial distributions. *Biometrika*, *53*, 167–182.

Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology, 76*(6), 1000–1027.

Sen, S. K., & Gan, S. K. (1983). A mathematical extension of the idea of bibliographic coupling and its applications. *Annals of Library and Information Studies, 30*(2), 78–82.

Smith, C. H. (2000). The classical music navigator. http://people.wku.edu/charles.smith/music/. Accessed 17 November 2014

Smith, C. H., & Georges, P. (2014). Composer similarities through 'The Classical Music Navigator': Similarity inference from composer influences. *Empirical Studies of the Arts, 32*(2), 205–229.

Spósito, M. B., Amorim, L., Ribeiro, P. J. Jr., Bassanezi, R. B., & Krainski, E. T. (2007). Spatial pattern of trees affected by black spot in citrus groves in Brazil. *Plant Disease, 91*(1), 36–40.

Wallet, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data*. Hoboken, NJ: Wiley.

Wiley, D. L. (1998). Cited references on the Web: A review of ISI's 'Web of Science'. *Searcher, 6*(1), 32–39, 57.

Wolfram, D. (2003). *Applied informetrics for information retrieval research*. Westport, CT: Libraries Unlimited.

Zhao, D., & Strotmann, A. (2014). In-text author citation analysis: Feasibility, benefits, and limitations. *Journal of the Association for Information Science and Technology, 65*(11), 2348–2358.