How to Normalize Co-Occurrence Data? An Analysis of Some Well-Known Similarity Measures

Nees Jan van Eck and Ludo Waltman

ERIM REPORT SERIES *RESEARCH IN MANAGEMENT*	
ERIM Report Series reference number	ERS-2009-001-LIS
Publication	January 2009
Number of pages	42
Persistent paper URL	http://hdl.handle.net/1765/14528
Email address corresponding author	nvaneck@ese.eur.nl
Address	Erasmus Research Institute of Management (ERIM) RSM Erasmus University / Erasmus School of Economics Erasmus Universiteit Rotterdam P.O.Box 1738 3000 DR Rotterdam, The Netherlands Phone: + 31 10 408 1182 Fax: + 31 10 408 9640 Email: info@erim.eur.nl Internet: www.erim.eur.nl

Bibliographic data and classifications of all the ERIM reports are also available on the ERIM website: www.erim.eur.nl

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

REPORT SERIES

RESEARCH IN MANAGEMENT

ABSTRACT AND KEYWORDS	
Abstract	In scientometric research, the use of co-occurrence data is very common. In many cases, a similarity measure is employed to normalize the data. However, there is no consensus among researchers on which similarity measure is most appropriate for normalization purposes. In this paper, we theoretically analyze the properties of similarity measures for co-occurrence data, focusing in particular on four well-known measures: the association strength, the cosine, the inclusion index, and the Jaccard index. We also study the behavior of these measures empirically. Our analysis reveals that there exist two fundamentally different types of similarity measures, namely set-theoretic measures and probabilistic measures. The association strength is a probabilistic measure, while the cosine, the inclusion index, and the Jaccard index are set-theoretic measures. Both our theoretical and our empirical results indicate that co-occurrence data can best be normalized using a probabilistic measure. This provides strong support for the use of the association strength in scientometric research.
Free Keywords	similarity measure, association strength, cosine, inclusion index, Jaccard index
Availability	
Classifications	

How to normalize co-occurrence data? An analysis of some well-known similarity measures

Nees Jan van Eck[*†] Ludo Waltman[*]

[*]Econometric Institute, Erasmus School of Economics
Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
E-mail: {nvaneck,lwaltman}@ese.eur.nl

[†]Centre for Science and Technology Studies, Leiden University
P.O. Box 905, 2300 AX Leiden, The Netherlands

Abstract

In scientometric research, the use of co-occurrence data is very common. In many cases, a similarity measure is employed to normalize the data. However, there is no consensus among researchers on which similarity measure is most appropriate for normalization purposes. In this paper, we theoretically analyze the properties of similarity measures for co-occurrence data, focusing in particular on four well-known measures: the association strength, the cosine, the inclusion index, and the Jaccard index. We also study the behavior of these measures empirically. Our analysis reveals that there exist two fundamentally different types of similarity measures, namely set-theoretic measures and probabilistic measures. The association strength is a probabilistic measure, while the cosine, the inclusion index, and the Jaccard index are set-theoretic measures. Both our theoretical and our empirical results indicate that co-occurrence data can best be normalized using a probabilistic measure. This provides strong support for the use of the association strength in scientometric research.

Keywords

Similarity measure, association strength, cosine, inclusion index, Jaccard index.

1 Introduction

The use of co-occurrence data is very common in scientometric research. Co-occurrence data can be used for a multitude of purposes. Co-citation data, for example, can be used to study relations among authors or journals, co-authorship data can be used to study scientific cooperation, and data on co-occurrences of words can be used to construct so-called co-word maps, which are maps that provide a visual representation of the structure of a scientific field. Usually, when co-occurrence data is used, a transformation is first applied to the data. The aim of such a transformation is to derive similarities from the data or, more specifically, to normalize the data. For example, when researchers study relations among authors based on co-citation data, they typically derive similarities from the data and then analyze these similarities using multivariate analysis techniques such as multidimensional scaling and hierarchical clustering (e.g., McCain, 1990; White and Griffith, 1981; White and McCain, 1998). Likewise, when researchers use co-authorship data to study scientific cooperation, they typically apply a normalization to the data and then base their analysis on the normalized data (e.g., Glänzel, 2001; Luukkonen, Persson, and Sivertsen, 1992; Luukkonen, Tijssen, Persson, and Sivertsen, 1993).

In this paper, our focus is methodological. We study various measures for deriving similarities from co-occurrence data. Basically, there are two approaches that can be taken to derive similarities from co-occurrence data. We refer to these approaches as the direct and the indirect approach, but the approaches are also known as the local and the global approach (Ahlgren, Jarneving, and Rousseau, 2003; Jarneving, 2008). Similarity measures that implement the direct approach are referred to as direct similarity measures in this paper, while similarity measures that implement the indirect approach are referred to as indirect similarity measures.

The indirect approach to derive similarities from co-occurrence data relies on co-occurrence profiles. The co-occurrence profile of an object is a vector that contains the number of co-occurrences of the object with each other object. Indirect similarity measures determine the similarity between two objects by comparing the co-occurrence profiles of the objects. The indirect approach is mainly used for co-citation data (e.g., McCain, 1990, 1991; White and Griffith, 1981; White and McCain, 1998). From a theoretical point of view, the approach is quite well understood (Ahlgren et al., 2003; Van Eck and Waltman, 2008).

In this paper, we focus most of our attention on the direct approach to derive similarities from co-occurrence data. Direct similarity measures determine the similarity between two ob-

jects by taking the number of co-occurrences of the objects and adjusting this number for the total number of occurrences or co-occurrences of each of the objects. Researchers use several different direct similarity measures. The cosine and the Jaccard index are especially popular, but other measures are also regularly used. However, relatively little is known about the theoretical properties of the various measures. Also, there is no consensus among researchers on which measure is most appropriate for a particular purpose. In this paper, we theoretically analyze some well-known direct similarity measures and we compare their properties. We also study the behavior of the measures empirically. Usually, when a direct similarity measure is applied to co-occurrence data, the purpose is to normalize the data, that is, to correct the data for differences in the total number of occurrences or co-occurrences of objects. The main question that we try to answer in this paper is therefore as follows: Which direct similarity measures are appropriate for normalizing co-occurrence data and which are not? Interestingly, despite their popularity, the cosine and the Jaccard index turn out not to be appropriate measures for normalization purposes. We argue that an appropriate measure for normalizing co-occurrence data is the association strength (Van Eck and Waltman, 2007; Van Eck, Waltman, van den Berg, and Kaymak, 2006), also referred to as the proximity index (e.g., Peters and van Raan, 1993a; Rip and Courtial, 1984) or the probabilistic affinity index (e.g., Zitt, Bassecoulard, and Okubo, 2000). Although this measure is somewhat less well-known, it turns out to have the right theoretical properties for normalizing co-occurrence data.

This paper is organized as follows. We first provide an overview of the most popular direct similarity measures. We then analyze these measures theoretically. We also look for empirical relations among the measures. Finally, we answer the question which direct similarity measures are appropriate for normalizing co-occurrence data and which are not.

2 Overview of direct similarity measures

In this section, we provide an overview of the most popular direct similarity measures. The overview is based on a survey of the scientometric literature.

We first introduce some mathematical notation. Let $\mathbf{O}$ denote an occurrence matrix of order $m \times n$. The columns of $\mathbf{O}$ correspond with the objects of which we want to analyze the co-occurrences. There are n such objects, denoted by $1, \ldots, n$. The objects can be, for exam-

ple, authors (e.g., White and McCain, 1998), countries (e.g., Glänzel, 2001; Zitt et al., 2000), documents (e.g., Gmür, 2003; Klavans and Boyack, 2006b), journals (e.g., Boyack, Klavans, and Börner, 2005; Klavans and Boyack, 2006a), Web pages (e.g., Vaughan, 2006; Vaughan and You, 2006), or words (e.g., Kopcsa and Schiebel, 1998). The rows of $\mathbf{O}$ usually correspond with documents. m then denotes the number of documents on which the co-occurrence analysis is based. Sometimes the rows of $\mathbf{O}$ do not correspond with documents. In Web co-link analysis, for example, the rows of $\mathbf{O}$ correspond with Web pages (e.g., Vaughan, 2006; Vaughan and You, 2006). Throughout this paper, however, we assume for simplicity that the rows of $\mathbf{O}$ always correspond with documents. Another assumption that we make is that $\mathbf{O}$ is a binary matrix, that is, each element of $\mathbf{O}$ equals either zero or one. Let o_{ki} denote the element in the kth row and ith column of $\mathbf{O}$. o_{ki} equals one if object i occurs in the document that corresponds with the kth row of $\mathbf{O}$, and it equals zero otherwise. Let $\mathbf{C}$ denote the co-occurrence matrix of the objects $1, \ldots, n$. $\mathbf{C}$ is a symmetric non-negative matrix of order $n \times n$. Let c_{ij} denote the element in the ith row and jth column of $\mathbf{C}$. For $i \neq j$, c_{ij} equals the number of co-occurrences of objects i and j. For $i = j$, c_{ij} equals the number of occurrences of object i. Clearly, for all i and j, $c_{ij} = \sum_{k=1}^{m} o_{ki} o_{kj}$. It follows from this that $\mathbf{C} = \mathbf{O}^{\mathrm{T}}\mathbf{O}$, where $\mathbf{O}^{\mathrm{T}}$ denotes the transpose of $\mathbf{O}$. Moreover, the assumption that $\mathbf{O}$ is a binary matrix implies that $\mathbf{C}$ is an integer matrix.

As we discussed in the introduction, there are two types of measures for determining similarities between objects based on co-occurrence data. We refer to these two types of measures as direct similarity measures and indirect similarity measures. Indirect similarity measures, also known as global similarity measures (Ahlgren et al., 2003; Jarneving, 2008), determine the similarity between two objects i and j by comparing the ith and the jth row (or column) of the co-occurrence matrix $\mathbf{C}$. The more similar the co-occurrence profiles in these two rows (or columns) of $\mathbf{C}$, the higher the similarity between i and j. Indirect similarity measures are especially popular for author co-citation analysis (e.g., McCain, 1990; White and Griffith, 1981; White and McCain, 1998) and journal co-citation analysis (e.g., McCain, 1991). We refer to Ahlgren et al. (2003) and Van Eck and Waltman (2008) for a detailed discussion of the properties of various indirect similarity measures. In this paper, we focus most of our attention on direct similarity measures, also known as local similarity measures (Ahlgren et al., 2003; Jarneving, 2008). Direct similarity measures determine the similarity between two objects i and j by taking the number of co-occurrences of i and j and adjusting this number for the

total number of occurrences or co-occurrences of i and the total number of occurrences or co-occurrences of j. We note that in some studies similarities between objects are determined by comparing columns of the occurrence matrix $\mathbf{O}$ (e.g., Leydesdorff and Vaughan, 2006; Schneider, Larsen, and Ingwersen, 2008). In most cases, this approach is mathematically equivalent to the use of a direct similarity measure.[1]

Let s_i denote either the total number of occurrences of object i or the total number of co-occurrences of object i. In the first case we have

$$s_i = c_{ii} = \sum_{k=1}^{m} o_{ki}, \tag{1}$$

while in the second case we have

$$s_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} c_{ij}. \tag{2}$$

Both definitions are used in scientometric research (see also Leydesdorff, 2008), but the first definition seems to be more popular. We now provide a formal definition of a direct similarity measure.

Definition 2.1. A *direct similarity measure* is defined as a function $S(c_{ij}, s_i, s_j)$ that has the following three properties:

- The domain of $S(c_{ij}, s_i, s_j)$ equals

$$\mathcal{D}_S = \big\{(c_{ij}, s_i, s_j) \in \mathbb{R}^3 \big| 0 \leq c_{ij} \leq \min(s_i, s_j) \text{ and } s_i, s_j > 0\big\}. \tag{3}$$

- The range of $S(c_{ij}, s_i, s_j)$ is a subset of $\mathbb{R}$.

- $S(c_{ij}, s_i, s_j)$ is symmetric in s_i and s_j, that is, $S(c_{ij}, s_i, s_j) = S(c_{ij}, s_j, s_i)$ for all $(c_{ij}, s_i, s_j) \in \mathcal{D}_S$.

Based on this definition, a number of observations can be made. First, the definition does not require that c_{ij}, s_i, and s_j have integer values. Allowing for non-integer values of c_{ij}, s_i, and s_j simplifies the mathematical analysis of direct similarity measures. Second, even though most

[1] Leydesdorff and Vaughan (2006) and Schneider et al. (2008) use the Pearson correlation to compare columns of the occurrence matrix $\mathbf{O}$. As shown by Guilford (1973), applying the Pearson correlation to a binary occurrence matrix is mathematically equivalent to applying the so-called phi coefficient to the corresponding co-occurrence matrix.

direct similarity measures take values between zero and one, the definition allows measures to have a different range. And third, because the definition requires direct similarity measures to be symmetric in s_i and s_j, it does not cover asymmetric similarity measures such as those discussed by Egghe and Michel (2002, 2003). As a final observation, we note that Definition 2.1 is quite general. More specific definitions for special classes of direct similarity measures will be provided later on in this paper. We now define the notion of monotonic relatedness of direct similarity measures.

Definition 2.2. Two direct similarity measures $S_1(c_{ij}, s_i, s_j)$ and $S_2(c_{ij}, s_i, s_j)$ are said to be *monotonically related* if and only if

$$S_1(c_{ij}, s_i, s_j) < S_1(c'_{ij}, s'_i, s'_j) \Leftrightarrow S_2(c_{ij}, s_i, s_j) < S_2(c'_{ij}, s'_i, s'_j) \tag{4}$$

for all $(c_{ij}, s_i, s_j), (c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$.

Monotonic relatedness of direct similarity measures is important because certain multivariate analysis techniques that are frequently used in scientometric research are insensitive to monotonic transformations of similarities. This is for example the case for ordinal or non-metric multidimensional scaling (e.g., Borg and Groenen, 2005) and for single linkage and complete linkage hierarchical clustering (e.g., Anderberg, 1973).

Based on a survey of the literature, we have identified the most popular direct similarity measures in the field of scientometrics. These measures are defined as

$$S_{\mathrm{A}}(c_{ij}, s_i, s_j) = \frac{c_{ij}}{s_i s_j}, \tag{5}$$

$$S_{\mathrm{C}}(c_{ij}, s_i, s_j) = \frac{c_{ij}}{\sqrt{s_i s_j}}, \tag{6}$$

$$S_{\mathrm{I}}(c_{ij}, s_i, s_j) = \frac{c_{ij}}{\min(s_i, s_j)}, \tag{7}$$

$$S_{\mathrm{J}}(c_{ij}, s_i, s_j) = \frac{c_{ij}}{s_i + s_j - c_{ij}}. \tag{8}$$

We refer to these measures as, respectively, the association strength, the cosine, the inclusion index, and the Jaccard index. Assuming that c_{ij} is an integer, each of the measures takes values between zero and one. Moreover, it is not difficult to see that the measures satisfy

$$S_{\mathrm{A}}(c_{ij}, s_i, s_j) \leq S_{\mathrm{J}}(c_{ij}, s_i, s_j) \leq S_{\mathrm{C}}(c_{ij}, s_i, s_j) \leq S_{\mathrm{I}}(c_{ij}, s_i, s_j). \tag{9}$$

We now discuss each of the measures.

The association strength defined in (5) is used by Van Eck and Waltman (2007) and Van Eck et al. (2006).[2] Under various names, the measure is also used in a number of other studies. Hinze (1994), Leclerc and Gagné (1994), Peters and van Raan (1993a), and Rip and Courtial (1984) refer to the measure as the proximity index, while Leydesdorff (2008) and Zitt et al. (2000) refer to it as the probabilistic affinity (or activity) index. The measure is also employed by Luukkonen et al. (1992, 1993), but in their work it does not have a name. The association strength is proportional to the ratio between on the one hand the observed number of co-occurrences of objects i and j and on the other hand the expected number of co-occurrences of objects i and j under the assumption that occurrences of i and j are statistically independent. We will come back to this interpretation later on in this paper. The association strength corresponds with the pseudo-cosine measure discussed by Jones and Furnas (1987) and is monotonically related to the (pointwise) mutual information measure used in the field of computational linguistics (e.g., Church and Hanks, 1990; Manning and Schütze, 1999). Measures equivalent to the association strength sometimes also appear outside the field of scientometrics (Cox and Cox, 2001, 2008; Hubálek, 1982).

The cosine defined in (6) equals the ratio between on the one hand the number of times that objects i and j are observed together and on the other hand the geometric mean of the number of times that object i is observed and the number of times that object j is observed. The measure can be interpreted as the cosine of the angle between the ith and the jth column of the occurrence matrix $\mathbf{O}$, where the columns of $\mathbf{O}$ are regarded as vectors in an m-dimensional space (e.g., Salton and McGill, 1983). The cosine seems to be the most popular direct similarity measure in the field of scientometrics. Frequently cited studies in which the measure is used include Braam, Moed, and van Raan (1991a,b), Klavans and Boyack (2006a), Leydesdorff (1989), Peters and van Raan (1993b), Peters, Braam, and van Raan (1995), Small (1994), Small and Sweeney (1985), and Small, Sweeney, and Greenlee (1985). The popularity of the cosine is largely due to the work of Salton in the field of information retrieval (e.g., Salton, 1963; Salton and McGill, 1983). The cosine is therefore sometimes referred to as Salton's measure (e.g., Glänzel, 2001; Glänzel, Schubert, and Czerwon, 1999; Luukkonen et al., 1993; Schubert

[2]The definition of the association strength used in these papers differs slightly from the definition provided in (5). However, since the two definitions are proportional to each other, the difference between them is not important. Throughout this section, direct similarity measures that are proportional to each other will simply be regarded as equivalent.

and Braun, 1990) or as the Salton index (e.g., Morillo, Bordons, and Gómez, 2003). In some studies, a measure called the equivalence index is used (e.g., Callon, Courtial, and Laville, 1991; Kostoff, Eberhart, and Toothman, 1999; Law and Whittaker, 1992; Palmer, 1999). This measure equals the square of the cosine. Outside the fields of scientometrics and information retrieval, the cosine is also known as the Ochiai coefficient (e.g., Cox and Cox, 2001, 2008; Hubálek, 1982; Sokal and Sneath, 1963).

Examples of the use of the inclusion index defined in (7) can be found in the work of Kostoff, del Río, Humenik, García, and Ramírez (2001), McCain (1995), Peters and van Raan (1993a), Rip and Courtial (1984), Tijssen (1992, 1993), and Tijssen and van Raan (1989). We note that a measure somewhat different from the one defined in (7) is sometimes also called the inclusion index (e.g., Braam et al., 1991a; Kostoff et al., 1999; Peters et al., 1995; Qin, 2000). In the field of information retrieval, the inclusion index is referred to as the overlap measure (e.g., Jones and Furnas, 1987; Rorvig, 1999; Salton and McGill, 1983). More in general, the inclusion index is sometimes called the Simpson coefficient (e.g., Cox and Cox, 2001, 2008; Hubálek, 1982).

The Jaccard index defined in (8) equals the ratio between on the one hand the number of times that objects i and j are observed together and on the other hand the number of times that at least one of the two objects is observed. Small uses the Jaccard index in his early work on co-citation analysis (e.g., Small, 1973, 1981; Small and Greenlee, 1980). Other work in which the Jaccard index is used includes Heimeriks, Hörlesberger, and van den Besselaar (2003), Kopcsa and Schiebel (1998), Peters and van Raan (1993a), Peters et al. (1995), Rip and Courtial (1984), Van Raan and Tijssen (1993), Vaughan (2006), and Vaughan and You (2006). As shown by Anderberg (1973), the Jaccard index is monotonically related to the Dice coefficient, which is a well-known measure in information retrieval (e.g., Jones and Furnas, 1987; Rorvig, 1999; Salton and McGill, 1983) and other fields (e.g., Cox and Cox, 2001, 2008; Hubálek, 1982; Sokal and Sneath, 1963).

We note that, in addition to the four direct similarity measures discussed above, many more direct similarity measures have been used in scientometric research. However, the above four measures are by far the most popular ones, and we therefore focus most of our attention on them in this paper. The relations among various direct similarity measures are summarized in Table 1.

In the field of scientometrics, a number of studies have been performed in which different

Table 1: Relations among various direct similarity measures.

Measure	Alternative names	Monotonically related measures
association strength	probabilistic affinity index	(pointwise) mutual information
	proximity index	
	pseudo-cosine	
cosine	Ochiai coefficient	equivalence index
	Salton's index/measure	
inclusion index	overlap measure	
	Simpson coefficient	
Jaccard index		Dice coefficient

direct similarity measures are compared with each other. Boyack et al. (2005), Gmür (2003), Klavans and Boyack (2006a), Leydesdorff (2008), Luukkonen et al. (1993), and Peters and van Raan (1993a) report results of empirical comparisons of different measures. Theoretical analyses of relations between different measures can be found in the work of Egghe (2008) and Hamers et al. (1989). Properties of various measures are also studied theoretically by Egghe and Rousseau (2006). An extensive discussion of the issue of comparing different measures is provided by Schneider and Borlund (2007a,b). Other work that might be of interest has been done in the field of information retrieval. In the information retrieval literature, empirical comparisons of different direct similarity measures are discussed by Chung and Lee (2001) and Rorvig (1999) and a theoretical comparison is presented by Jones and Furnas (1987).[3] We further note that general overviews of a large number of direct similarity measures and their properties can be found in the statistical literature (Anderberg, 1973; Cox and Cox, 2001, 2008; Gower, 1985; Gower and Legendre, 1986) and also in the biological literature (Hubálek, 1982; Sokal and Sneath, 1963).

[3]The results reported by Jones and Furnas are probably not very relevant to scientometric research. This is because Jones and Furnas focus on the effect of term weights on similarity measures. In scientometric research, there is no natural analogue to the term weights used in information retrieval. The reason for this is that the occurrence matrices used in scientometric research contain elements that are usually restricted to zero and one, while the document-term matrices used in information retrieval contain term weights that often do not have this restriction.

3 Set-theoretic similarity measures

In this section and in the next one, we are concerned with two special classes of direct similarity measures. We discuss the class of set-theoretic similarity measures in this section and the class of probabilistic similarity measures in the next section. It turns out that there is a fundamental difference between the cosine, the inclusion index, and the Jaccard index on the one hand and the association strength on the other hand. The first three measures all belong to the class of set-theoretic similarity measures, while the last measure belongs to the class of probabilistic similarity measures. We assume from now on that s_i denotes the total number of occurrences of object i, that is, we assume that the definition of s_i in (1) is adopted. From a theoretical point of view, this definition is more convenient than the definition of s_i in (2). We note that proofs of the theoretical results that we present in this section and in the next one are provided in the appendix.

Each column of an occurrence matrix can be seen as a representation of a set, namely the set of all documents in which a certain object occurs (cf Egghe and Rousseau, 2006). Consequently, a natural approach to determine the similarity between two objects i and j seems to be to determine the similarity between on the one hand the set of all documents in which i occurs and on the other hand the set of all documents in which j occurs. We refer to direct similarity measures that take this approach as set-theoretic similarity measures. In other words, set-theoretic similarity measures are direct similarity measures that are based on the notion of similarity between sets. In this section, we theoretically analyze the properties of set-theoretic similarity measures. We note that these properties are also studied theoretically by Baulieu (1989, 1997), Egghe and Michel (2002, 2003), Egghe and Rousseau (2006), and Janson and Vegelius (1981).

There are a number of properties of which we believe that it is natural to expect that any set-theoretic similarity measure $S(c_{ij}, s_i, s_j)$ has them. Three of these properties are given below.

Property 3.1. If $c_{ij} = 0$, then $S(c_{ij}, s_i, s_j)$ takes its minimum value.

Property 3.2. For all $\alpha > 0$, $S(\alpha c_{ij}, \alpha s_i, \alpha s_j) = S(c_{ij}, s_i, s_j)$.

Property 3.3. If $s'_i > s_i$ and $c_{ij} > 0$, then $S(c_{ij}, s'_i, s_j) < S(c_{ij}, s_i, s_j)$.

Property 3.1 is based on the idea that the similarity between two sets should be minimal if the sets are disjoint, that is, if they have no elements in common. Property 3.2 is based on the

idea that the similarity between two sets should remain unchanged in the case of a proportional increase or decrease in both the number of elements of each of the sets and the number of elements of the intersection of the sets. Egghe and Rousseau (2006) refer to this idea as replication invariance. It underlies the notion of Lorenz similarity that is studied by Egghe and Rousseau. A similar idea is also used by Janson and Vegelius (1981), who call it homogeneity. Property 3.3 is based on the idea that the similarity between two sets should decrease if an element is added to one of the sets and this element does not belong to the other set. A similar idea is used by Baulieu (1989, 1997). It is not difficult to see that Properties 3.1, 3.2, and 3.3 are independent of each other, that is, none of the properties is implied by the others. We regard Properties 3.1, 3.2, and 3.3 as the characterizing properties of set-theoretic similarity measures. This is formally stated in the following definition.

Definition 3.1. A *set-theoretic similarity measure* is defined as a direct similarity measure $S(c_{ij}, s_i, s_j)$ that has Properties 3.1, 3.2, and 3.3.

This definition implies that the cosine defined in (6) and the Jaccard index defined in (8) are set-theoretic similarity measures. The association strength defined in (5) does not have Property 3.2 and is therefore not a set-theoretic similarity measure. The inclusion index defined in (7) is also not a set-theoretic similarity measure. This is because the inclusion index does not have Property 3.3. However, the inclusion index does have the following property, which is a weakened version of Property 3.3.

Property 3.4. If $s_i' > s_i$ and $c_{ij} > 0$, then $S(c_{ij}, s_i', s_j) \leq S(c_{ij}, s_i, s_j)$.

This property naturally leads to the following definition.

Definition 3.2. A *weak set-theoretic similarity measure* is defined as a direct similarity measure $S(c_{ij}, s_i, s_j)$ that has Properties 3.1, 3.2, and 3.4.

It follows from this definition that the inclusion index is a weak set-theoretic similarity measure. We note that our definition of a set-theoretic similarity measure seems to be more restrictive than the definition of a Lorenz similarity function that is provided by Egghe and Rousseau (2006). This is because a Lorenz similarity function need not have Properties 3.1 and 3.3.

In addition to Properties 3.1, 3.2, and 3.3, there are some other properties that we consider indispensable for any set-theoretic similarity measure $S(c_{ij}, s_i, s_j)$. Four of these properties are given below.

Property 3.5. If $S(c_{ij}, s_i, s_j)$ takes its minimum value, then $c_{ij} = 0$.

Property 3.6. If $c_{ij} = s_i = s_j$, then $S(c_{ij}, s_i, s_j)$ takes its maximum value.

Property 3.7. If $S(c_{ij}, s_i, s_j)$ takes its maximum value, then $c_{ij} = s_i = s_j$.

Property 3.8. For all $\alpha > 0$, if $c_{ij} < s_i$ or $c_{ij} < s_j$, then $S(c_{ij}+\alpha, s_i+\alpha, s_j+\alpha) > S(c_{ij}, s_i, s_j)$.

Properties 3.5, 3.6, and 3.7 are based on the idea that the similarity between two sets should be minimal only if the sets are disjoint and that it should be maximal if and only if the sets are equal. Property 3.8 is based on the idea that the similarity between two sets should increase if the same element is added to both sets. It turns out that Properties 3.5, 3.6, 3.7, and 3.8 are implied by Properties 3.1, 3.2, and 3.3. This is stated by the following proposition.

Proposition 3.1. *All set-theoretic similarity measures $S(c_{ij}, s_i, s_j)$ have Properties 3.5, 3.6, 3.7, and 3.8.*

We note that weak set-theoretic similarity measures need not have Properties 3.5, 3.7, and 3.8. They do have Property 3.6.

We now consider the following two properties.

Property 3.9. If $s'_i s'_j > s_i s_j$ and $c_{ij} > 0$, then $S(c_{ij}, s'_i, s'_j) < S(c_{ij}, s_i, s_j)$. If $s'_i s'_j = s_i s_j$, then $S(c_{ij}, s'_i, s'_j) = S(c_{ij}, s_i, s_j)$.

Property 3.10. If $s'_i + s'_j > s_i + s_j$ and $c_{ij} > 0$, then $S(c_{ij}, s'_i, s'_j) < S(c_{ij}, s_i, s_j)$. If $s'_i + s'_j = s_i + s_j$, then $S(c_{ij}, s'_i, s'_j) = S(c_{ij}, s_i, s_j)$.

It is easy to see that these properties both imply Property 3.3. Hence, Properties 3.9 and 3.10 are both stronger than Property 3.3. It can further be seen that the cosine has Property 3.9 and that the Jaccard index has Property 3.10. The following two propositions indicate the importance of Properties 3.9 and 3.10.

Proposition 3.2. *All set-theoretic similarity measures $S(c_{ij}, s_i, s_j)$ that have Property 3.9 are monotonically related to the cosine defined in* (6).

Proposition 3.3. *All set-theoretic similarity measures $S(c_{ij}, s_i, s_j)$ that have Property 3.10 are monotonically related to the Jaccard index defined in* (8).

Table 2: Summary of the properties of a number of direct similarity measures. If a measure has a certain property, this is indicated using a $\times$ symbol.

	Property												
	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	3.10	3.11	4.1	4.2
Association strength	$\times$		$\times$	$\times$	$\times$		$\times$		$\times$			$\times$	$\times$
Cosine	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$			$\times$	
Inclusion index	$\times$	$\times$		$\times$	$\times$	$\times$					$\times$	$\times$	
Jaccard index	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$		$\times$			

It follows from Proposition 3.2 that Properties 3.1, 3.2, and 3.9 characterize the class of all set-theoretic similarity measures that are monotonically related to the cosine. Likewise, it follows from Proposition 3.3 that Properties 3.1, 3.2, and 3.10 characterize the class of all set-theoretic similarity measures that are monotonically related to the Jaccard index. We now apply a similar idea to the inclusion index. The inclusion index has the following property.

Property 3.11. If $\min(s'_i, s'_j) > \min(s_i, s_j)$ and $c_{ij} > 0$, then $S(c_{ij}, s'_i, s'_j) < S(c_{ij}, s_i, s_j)$. If $\min(s'_i, s'_j) = \min(s_i, s_j)$, then $S(c_{ij}, s'_i, s'_j) = S(c_{ij}, s_i, s_j)$.

This property implies Property 3.4. Together with Properties 3.1 and 3.2, Property 3.11 characterizes the class of all weak set-theoretic similarity measures that are monotonically related to the inclusion index. This is an immediate consequence of the following proposition.

Proposition 3.4. *All weak set-theoretic similarity measures $S(c_{ij}, s_i, s_j)$ that have Property 3.11 are monotonically related to the inclusion index defined in* (7)*.*

In the above discussion, we have introduced a large number of properties that a direct similarity measure may or may not have. For convenience, in Table 2 we summarize for the association strength, the cosine, the inclusion index, and the Jaccard index which of these properties they have and which they do not have. We note that the last two properties in the table will be introduced in the next section.

In order to provide some additional insight into the relations among various (weak and non-weak) set-theoretic similarity measures, we now introduce what we call the generalized similarity index (for a similar idea, see Warrens, 2008).

Definition 3.3. The *generalized similarity index* is defined as a direct similarity measure that is given by

$$S_{\mathrm{G}}(c_{ij}, s_i, s_j; p) = \frac{2^{1/p} c_{ij}}{(s_i^p + s_j^p)^{1/p}}, \tag{10}$$

where p denotes a parameter that takes values in $\mathbb{R} \setminus \{0\}$.

For all values of the parameter p, the generalized similarity index takes values between zero and one. The index equals the ratio between on the one hand the number of times that objects i and j are observed together and on the other hand a power mean of the number of times that object i is observed and the number of times that object j is observed. (Power means, also known as generalized means or Hölder means, are a generalization of arithmetic, geometric, and harmonic means.) An interesting property of the generalized similarity index is that, for various values of p, the index reduces to a well-known (weak or non-weak) set-theoretic similarity measure. More specifically, it can be seen that

$$\lim_{p \to -\infty} S_{\mathrm{G}}(c_{ij}, s_i, s_j; p) = \frac{c_{ij}}{\min(s_i, s_j)}, \tag{11}$$

$$S_{\mathrm{G}}(c_{ij}, s_i, s_j; -1) = \frac{1}{2}\left(\frac{c_{ij}}{s_i} + \frac{c_{ij}}{s_j}\right), \tag{12}$$

$$\lim_{p \to 0} S_{\mathrm{G}}(c_{ij}, s_i, s_j; p) = \frac{c_{ij}}{\sqrt{s_i s_j}}, \tag{13}$$

$$S_{\mathrm{G}}(c_{ij}, s_i, s_j; 1) = \frac{2c_{ij}}{s_i + s_j}, \tag{14}$$

$$S_{\mathrm{G}}(c_{ij}, s_i, s_j; 2) = \frac{\sqrt{2} c_{ij}}{\sqrt{s_i^2 + s_j^2}}, \tag{15}$$

$$\lim_{p \to \infty} S_{\mathrm{G}}(c_{ij}, s_i, s_j; p) = \frac{c_{ij}}{\max(s_i, s_j)}, \tag{16}$$

where (11), (13), and (16) follow from the properties of power means as discussed by, for example, Hardy, Littlewood, and Pólya (1952). Equations (11) and (12) indicate that for $p \to -\infty$ the generalized similarity index equals the inclusion index and that for $p = -1$ it equals the so-called joint conditional probability measure that is used by McCain (1995). The latter measure is more generally known as one of the Kulczynski coefficients (e.g., Cox and Cox, 2001, 2008; Hubálek, 1982; Sokal and Sneath, 1963). It is easy to see that this measure is a set-theoretic similarity measure.[4] Equations (13) and (14) indicate that for $p \to 0$ the generalized

[4]This contrasts with Janson and Vegelius (1981), who argue that the measure in (12) does not have completely satisfactory properties.

similarity index equals the cosine and that for $p = 1$ it equals the Dice coefficient. It follows from (8) and (14) that

$$S_{\mathrm{G}}(c_{ij}, s_i, s_j; 1) = \frac{2S_{\mathrm{J}}(c_{ij}, s_i, s_j)}{S_{\mathrm{J}}(c_{ij}, s_i, s_j) + 1}, \tag{17}$$

which implies that for $p = 1$ the generalized similarity index is monotonically related to the Jaccard index. Equations (15) and (16) indicate that for $p = 2$ and $p \to \infty$ the generalized similarity index equals, respectively, the measures N and O_2 that are studied by Egghe and Michel (2002, 2003) and Egghe and Rousseau (2006). It is clear that N is a set-theoretic similarity measure and that O_2 is a weak set-theoretic similarity measure. Measures equivalent to (16) are also discussed by Cox and Cox (2001, 2008) and Hubálek (1982).

The following proposition points out an important property of the generalized similarity index.

Proposition 3.5. *For all finite values of the parameter p, the generalized similarity index defined in* (10) *is a set-theoretic similarity measure.*

This proposition states that the generalized similarity index describes an entire class of set-theoretic similarity measures. Each member of this class corresponds with a particular value of p. Only in the limit case in which $p \to \pm\infty$, the generalized similarity index is not a set-theoretic similarity measure. In this limit case, the generalized similarity index is a weak set-theoretic similarity measure.

4 Probabilistic similarity measures

In the previous section, we discussed the class of set-theoretic similarity measures. The cosine, the inclusion index, and the Jaccard index turned out to be (weak or non-weak) set-theoretic similarity measures. The association strength, however, turned out not to belong to the class of set-theoretic similarity measures. In this section, we discuss the class of probabilistic similarity measures. This is the class to which the association strength turns out to belong.

We are interested in direct similarity measures $S(c_{ij}, s_i, s_j)$ that have the following two properties.

Property 4.1. If $s_1 = s_2 = \ldots = s_n$, then $S(c_{ij}, s_i, s_j) = \alpha c_{ij}$ for all $i \neq j$ and for some $\alpha > 0$.

Property 4.2. For all $\alpha > 0$, $S(\alpha c_{ij}, \alpha s_i, s_j) = S(c_{ij}, s_i, s_j)$.

Property 4.1 requires that, if all objects occur equally frequently, the similarity between two objects is proportional to the number of co-occurrences of the objects. Property 4.2 requires that the similarity between two objects remains unchanged in the case of a proportional increase or decrease in on the one hand the number of co-occurrences of the objects and on the other hand the number of occurrences of one of the objects. (Notice the difference between this property and Property 3.2.) We regard Properties 4.1 and 4.2 as the characterizing properties of probabilistic similarity measures. This results in the following definition.

Definition 4.1. A *probabilistic similarity measure* is defined as a direct similarity measure $S(c_{ij}, s_i, s_j)$ that has Properties 4.1 and 4.2.

The cosine, the inclusion index, and the Jaccard index do not have Property 4.2 and therefore are not probabilistic similarity measures. The association strength, on the other hand, is a probabilistic similarity measure, since it has both Property 4.1 and Property 4.2. In this respect, the association strength is quite unique, as the following proposition indicates.

Proposition 4.1. *All probabilistic similarity measures are proportional to the association strength defined in* (5)*.*

This proposition states that the class of probabilistic similarity measures consists only of the association strength and of measures that are proportional to the association strength. There are no other measures that belong to the class of probabilistic similarity measures. The following result is an immediate consequence of Proposition 4.1.

Corollary 4.2. *A direct similarity measure cannot be both a (weak or non-weak) set-theoretic similarity measure and a probabilistic similarity measure.*

This result makes clear that there is a fundamental difference between set-theoretic similarity measures and probabilistic similarity measures. In other words, there is a fundamental difference between measures such as the cosine, the inclusion index, and the Jaccard index on the one hand and the association strength on the other hand. We will come back to this difference later on in this paper.

We now explain the rationale for Properties 4.1 and 4.2. To do so, we first discuss why direct similarity measures are applied to co-occurrence data. The number of co-occurrences of

two objects can be seen as the result of two independent effects. We refer to these effects as the similarity effect and the size effect.[5] The similarity effect is the effect that, other things being equal, more similar objects have more co-occurrences. The size effect is the effect that, other things being equal, an object that occurs more frequently has more co-occurrences with other objects. If one is interested in the similarity between two objects, the number of co-occurrences of the objects is in general not an appropriate measure. This is because, due to the size effect, the number of co-occurrences is likely to give a distorted picture of the similarity between the objects (see also Waltman and van Eck, 2007). Two frequently occurring objects, for example, may have a large number of co-occurrences and may therefore look very similar. However, it is quite well possible that the large number of co-occurrences of the objects is completely due to their high frequency of occurrence (i.e., the size effect) and has nothing to do with their similarity. Usually, when a direct similarity measure is applied to co-occurrence data, the aim is to correct the data for the size effect.

Based on the above discussion, the idea underlying Property 4.1 can be explained as follows. Property 4.1 is concerned with the behavior of a direct similarity measure in the special case in which all objects occur equally frequently. In this special case, the size effect is equally strong for all objects, which means that, unlike in the more general case, the number of co-occurrences of two objects is an appropriate measure of the similarity between the objects. Taking this into account, it is natural to expect that in the special case considered by Property 4.1 a direct similarity measure does not transform the co-occurrence frequencies of objects in any significant way. Property 4.1 implements this idea by requiring that, if all objects occur equally frequently, the similarity between two objects is proportional to the number of co-occurrences of the objects.

We now consider Property 4.2. The idea underlying this property can best be clarified by means of an example. Consider an arbitrary object i, and suppose that the total number of occurrences of i doubles. It can then be expected that the total number of co-occurrences of i also doubles, at least approximately. Suppose that the total number of co-occurrences of i indeed doubles and that the new co-occurrences of i are distributed over the other objects in the same way as the old co-occurrences of i. This simply means that the number of co-

[5]The similarity effect and the size effect can be seen as analogous to what statisticians call, respectively, interaction effects and main effects.

occurrences of i with each other object doubles. We believe that this increase in the number of occurrences and co-occurrences of i should not have any influence on the similarities between i and the other objects. This is because the number of occurrences of i and the number of co-occurrences of i with each other object have all increased proportionally, namely by a factor of two. Hence, relatively speaking, the frequency with which i co-occurs with each other object has not changed. This means that the increase in the number of co-occurrences of i with each other object is completely due to the size effect and has not been caused by the similarity effect. Taking this into account, it is natural to expect that the similarities between i and the other objects remain unchanged. Property 4.2 implements this idea. It does so not only for the case in which the number of occurrences and co-occurrences of an object doubles but more generally for any proportional increase or decrease in the number of occurrences and co-occurrences of an object. We note that the idea underlying Property 4.2 is not new. Ahlgren et al. (2003) and Van Eck and Waltman (2008) study properties of indirect similarity measures. A property that turns out to be particularly important is the so-called property of coordinate-wise scale invariance. Interestingly, this property relies on exactly the same idea as Property 4.2. Hence, direct similarity measures that have Property 4.2 and indirect similarity measures that have the property of coordinate-wise scale invariance are based on similar principles.

Finally, we discuss the probabilistic interpretation of probabilistic similarity measures (see also Leclerc and Gagné, 1994; Luukkonen et al., 1992, 1993; Zitt et al., 2000). Let p_i denote the probability that object i occurs in a randomly chosen document. It is clear that $p_i = s_i/m$. If two objects i and j occur independently of each other, the probability that they co-occur in a randomly chosen document equals $p_{ij} = p_i p_j$. The expected number of co-occurrences of i and j then equals $e_{ij} = mp_{ij} = mp_i p_j = s_i s_j/m$. A natural way to measure the similarity between i and j is to calculate the ratio between on the one hand the observed number of co-occurrences of i and j and on the other hand the expected number of co-occurrences of i and j under the assumption that i and j occur independently of each other (for a similar argument in a more general context, see De Solla Price, 1981). This results in a measure that equals c_{ij}/e_{ij}. This measure has a straightforward probabilistic interpretation. If $c_{ij}/e_{ij} > 1$, i and j co-occur more frequently than would be expected by chance. If, on the other hand, $c_{ij}/e_{ij} < 1$, i and j co-occur less frequently than would be expected by chance. It is easy to see that $c_{ij}/e_{ij} = mS_{\mathrm{A}}(c_{ij}, s_i, s_j)$. Hence, the measure c_{ij}/e_{ij} is proportional to the association strength

and, consequently, belongs to the class of probabilistic similarity measures. Since probabilistic similarity measures are all proportional to each other (this follows from Proposition 4.1), they all have a similar probabilistic interpretation as the measure c_{ij}/e_{ij}.

5 Empirical comparison

In the previous two sections, the differences between a number of well-known direct similarity measures were analyzed theoretically. It turned out that some measures have fundamentally different properties than others. An obvious question now is whether in practical applications there is much difference between the various measures. This is the question with which we are concerned in this section.

Leydesdorff (2008) reports the results of an empirical comparison of a number of direct and indirect similarity measures (for a theoretical explanation for some of the results, see Egghe, 2008). The measures are applied to a data set consisting of the co-citation frequencies of 24 authors, 12 from the field of information retrieval and 12 from the field of scientometrics.[6] It turns out that the direct similarity measures are strongly correlated with each other. The Spearman rank correlations between the association strength (referred to as the probabilistic affinity or activity index), the cosine, and the Jaccard index are all above 0.98. Hence, for the particular data set studied by Leydesdorff, there does not seem to be much difference between various direct similarity measures.

In this section, we examine whether the results reported by Leydesdorff hold more generally. To do so, we study three data sets, one consisting of co-citation frequencies of authors, one consisting of co-citation frequencies of journals, and one consisting of co-occurrence frequencies of terms. We refer to these data sets as, respectively, the author data set, the journal data set, and the term data set. The author data set consists of the co-citation frequencies of 100 authors in the field of information science in the period 1988–1995. The data set is studied extensively in a well-known paper by White and McCain (1998), and it is also used in one of our earlier papers (Van Eck and Waltman, 2008). The journal data set has not been studied before. The data set consists of the co-citation frequencies of 389 journals belonging to at least one of the following five subject categories of Thomson Reuters: *Business*, *Business-Finance*, *Economics*,

[6] The same data set is also studied by Ahlgren et al. (2003), Leydesdorff and Vaughan (2006), and Waltman and van Eck (2007).

Table 3: Main characteristics of the author data set, the journal data set, and the term data set.

	Author data set	Journal data set	Term data set
# objects	100	389	332
# documents	5 463	24 106	6 235
# occurrences	7 768	32 697	26 211
# co-occurrences	22 520	13 378	60 640
% zeros in co-occurrence matrix	26%	93%	74%

Management, and *Operations Research & Management Science*. The co-citation frequencies of the journals were determined based on citations in articles published between 2005 and 2007 to articles published in 2005. The term data set consists of the co-occurrence frequencies of 332 terms in the field of computational intelligence in the period 1996–2000. Co-occurrences of terms were counted in abstracts of articles published in important journals and conference proceedings in the computational intelligence field. For a more detailed description of the term data set, we refer to an earlier paper (Van Eck and Waltman, 2007). In Table 3, we summarize the main characteristics of the three data sets that we study.

In order to examine how the association strength, the cosine, the inclusion index, and the Jaccard index are empirically related to each other, we analyzed each of the three data sets as follows. We first calculated for each combination of two objects the value of each of the four similarity measures. For each combination of two similarity measures, we then drew a scatter plot that shows how the values of the two measures are related to each other. The scatter plots obtained for the author data set and the term data set are shown in Figures 1 and 2, respectively. The scatter plots obtained for the journal data set look very similar to the ones obtained for the term data set and are therefore not shown. After drawing the scatter plots, we determined for each combination of two similarity measures how strongly the values of the measures are correlated with each other. We calculated both the Pearson correlation and the Spearman correlation. The Pearson correlation was used to measure the degree to which the values of two measures are linearly related, while the Spearman correlation was used to measure the degree to which the values of two measures are monotonically related. When calculating the Pearson and Spearman correlations between the values of two measures, we only took into

account values above zero.[7] The correlations obtained for the three data sets are reported in Tables 4, 5, and 6. In each table, the values in the upper right part are Pearson correlations while the values in the lower left part are Spearman correlations.

The scatter plots in Figures 1 and 2 clearly show that in practical applications there can be substantial differences between different direct similarity measures. This is confirmed by the correlations in Tables 4, 5, and 6. These results differ from the ones reported by Leydesdorff (2008), who finds no substantial differences between different direct similarity measures. The difference between our results and the results of Leydesdorff is probably due to the unusual nature of the data set studied in Leydesdorff, in particular the small number of objects in the data set (24 authors) and the division of the objects into two strongly separated groups (the information retrieval researchers and the scientometricians). When looking in more detail at the scatter plots in Figures 1 and 2, it can be seen that the similarity measures that are strongest related to each other are the cosine and the Jaccard index. The same observation can be made in Tables 4, 5, and 6. The relatively strong relation between the cosine and the Jaccard index has been observed before and is discussed by Egghe (2008), Hamers et al. (1989), and Leydesdorff (2008). Apart from the relation between the cosine and the Jaccard index, the relations between the different similarity measures are quite weak. This is especially the case for the relations between the association strength and the other three measures. Consider, for example, how the association strength and the inclusion index are related to each other in the term data set. As can be seen in Figure 2, a low value of the association strength sometimes corresponds with a high value of the inclusion index and, the other way around, a low value of the inclusion index sometimes corresponds with a high value of the association strength. This clearly indicates that the relation between the two measures is rather weak, which is confirmed by the correlations in Table 6. It is further interesting to compare our empirical results with the theoretical results presented by Egghe (2008). Egghe mathematically studies relations between various (weak and non-weak) set-theoretic similarity measures under the simplifying assumption that the ratio of

[7] If two objects have zero co-occurrences, all four similarity measures have a value of zero. Co-occurrence matrices usually contain a large number of zeros (see Table 3). This leads to high correlations (close to one) between the values of the four similarity measures. We regard these high correlations as problematic because they do not properly reflect how the similarity measures are related to each other in the case of objects with a non-zero number of co-occurrences. To avoid the problem of the high correlations, we only took into account values above zero when calculating correlations between the values of the four similarity measures.

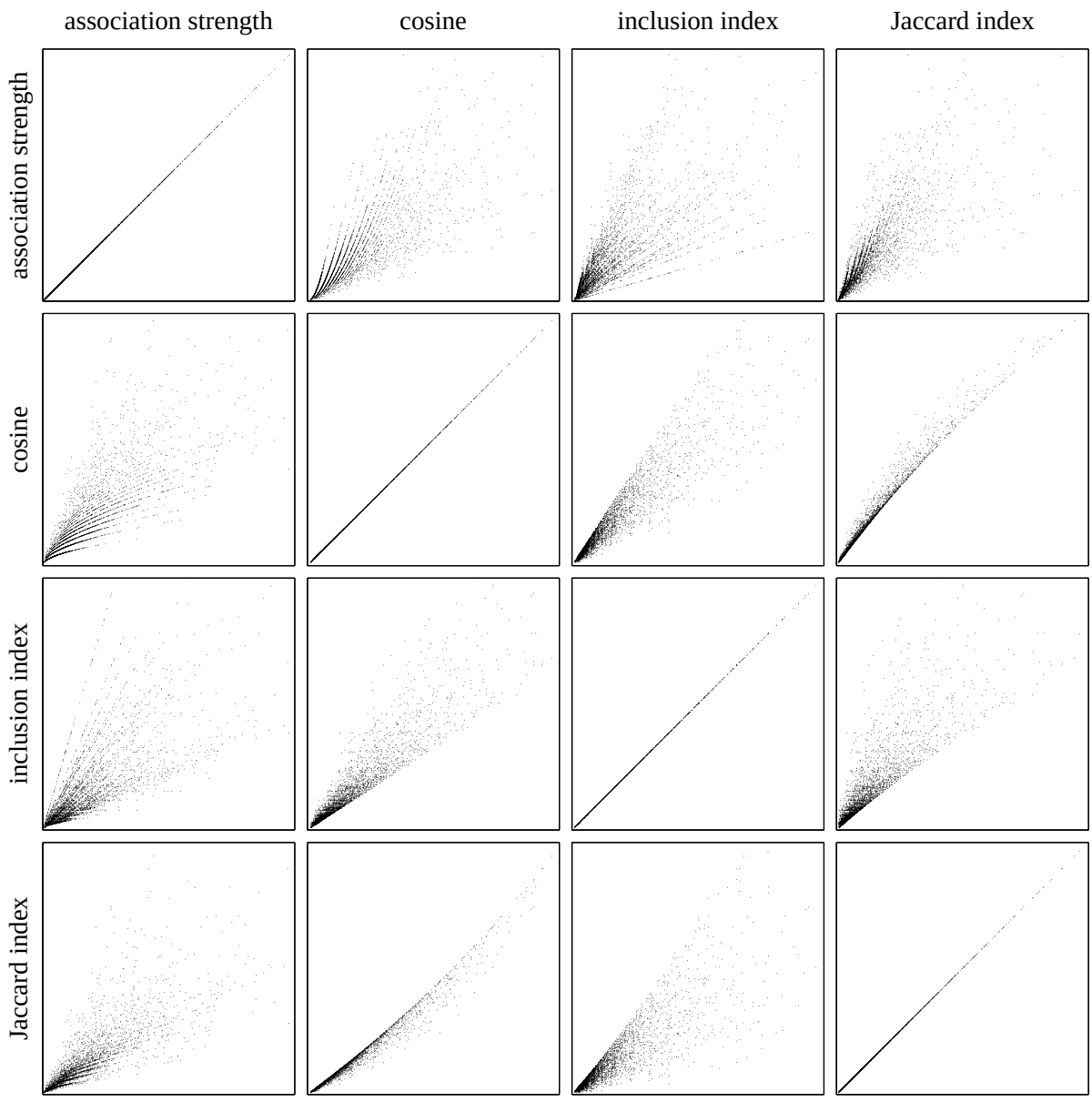

Figure 1: Scatter plots obtained for the author data set. In each plot, the lower left corner corresponds with the origin. The scales used for the different similarity measures are not the same.

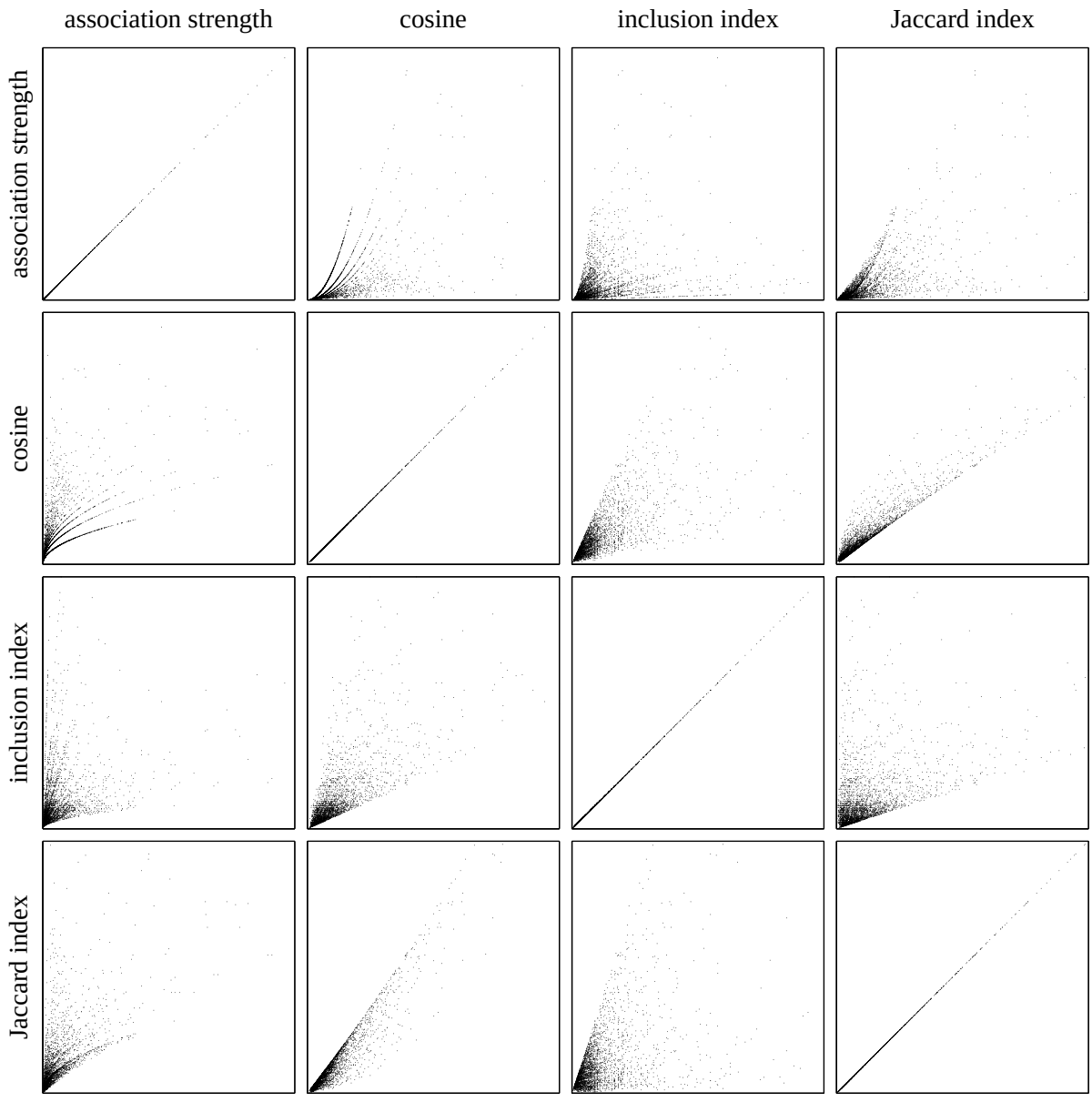

Figure 2: Scatter plots obtained for the term data set. In each plot, the lower left corner corresponds with the origin. The scales used for the different similarity measures are not the same.

Table 4: Correlations obtained for the author data set.

	Association strength	Cosine	Inclusion index	Jaccard index
Association strength		0.824	0.721	0.823
Cosine	0.913		0.929	0.987
Inclusion index	0.847	0.964		0.866
Jaccard index	0.920	0.994	0.931	

Table 5: Correlations obtained for the journal data set.

	Association strength	Cosine	Inclusion index	Jaccard index
Association strength		0.602	0.556	0.554
Cosine	0.892		0.800	0.971
Inclusion index	0.808	0.881		0.644
Jaccard index	0.832	0.952	0.708	

Table 6: Correlations obtained for the term data set.

	Association strength	Cosine	Inclusion index	Jaccard index
Association strength		0.653	0.347	0.688
Cosine	0.786		0.736	0.950
Inclusion index	0.562	0.799		0.511
Jaccard index	0.776	0.916	0.520	

the number of occurrences of two objects is fixed. He proves that, under this assumption, there exist simple monotonic (often linear) relations between many measures. However, especially for the inclusion index, the scatter plots in Figures 1 and 2 do not show such relations. Our empirical results therefore seem to indicate that the practical relevance of the theoretical results presented by Egghe might be somewhat limited.

The general conclusion that can be drawn from our empirical analysis is that there are quite significant differences between various direct similarity measures and, hence, that in practical applications it is important to use the measure that is most appropriate for ones purposes. In the next section, we discuss how an appropriate similarity measure can be chosen based on sound theoretical considerations. We focus in particular on the case in which a similarity measure is used for normalization purposes.

6 How to normalize co-occurrence data?

As we discussed in the previous sections, there are various ways in which similarities between objects can be determined based on co-occurrence data. The different types of similarity measures that can be used are shown in Figure 3. The first decision that one has to make is whether to use a direct or an indirect similarity measure. If one decides to use a direct similarity measure, one then has to decide whether to use a probabilistic or a set-theoretic similarity measure.

We first briefly discuss the use of indirect similarity measures. As pointed out by Schneider and Borlund (2007a), from a statistical perspective the use of an indirect similarity measure is a quite unconventional approach.[8] However, despite being unconventional, we do not believe that the approach has any fundamental statistical problems. Appropriate indirect similarity measures include the Bhattacharyya distance, the cosine,[9] and the Jensen-Shannon distance. These measures are known to have good theoretical properties (Van Eck and Waltman, 2008). A very popular indirect similarity measure, especially for author co-citation analysis (e.g., Mc-

[8] A similar approach is sometimes taken in psychological research (e.g., Rosenberg and Jones, 1972; Rosenberg, Nelson, and Vivekananthan, 1968). In the psychological literature, there is some discussion about the advantages and disadvantages of this approach (Drasgow and Jones, 1979; Simmen, 1996; Van der Kloot and van Herk, 1991).

[9] There are two different similarity measures, a direct and an indirect one, that are both referred to as the cosine. Here we mean the cosine as discussed by, for example, Ahlgren et al. (2003) and Van Eck and Waltman (2008). This is a different measure than the one defined in (6).

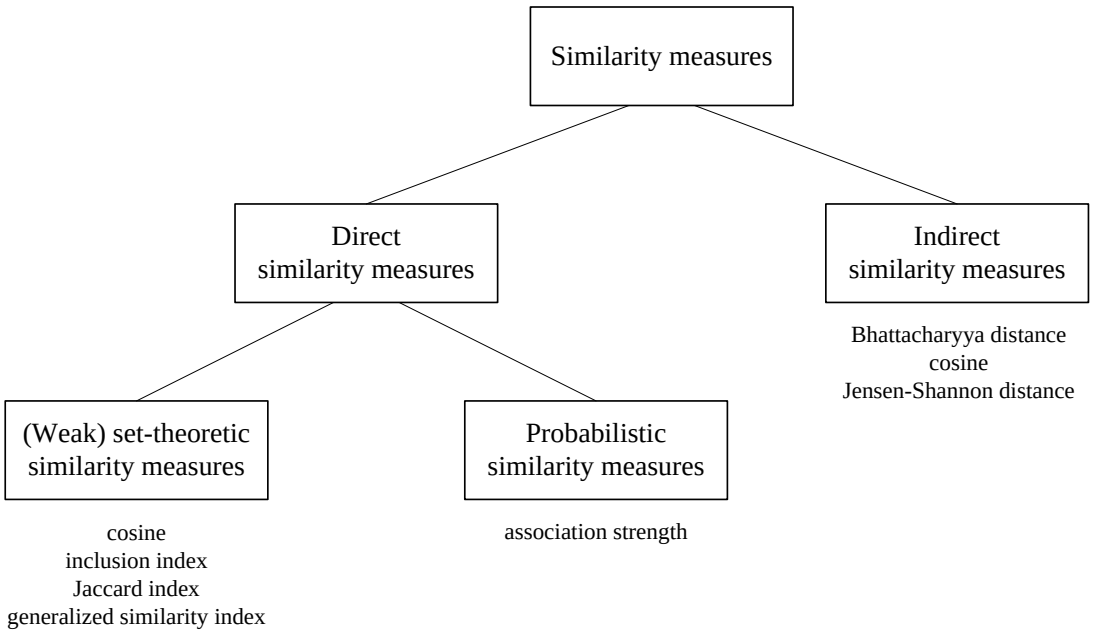

Figure 3: Different types of similarity measures.

Cain, 1990; White and Griffith, 1981; White and McCain, 1998), is the Pearson correlation. However, this measure does not have good theoretical properties and should therefore not be used (Ahlgren et al., 2003; Van Eck and Waltman, 2008). The chi-squared distance, which is proposed as an indirect similarity measure by Ahlgren et al. (2003), also does not have all the theoretical properties that we believe an appropriate indirect similarity measure should have (Van Eck and Waltman, 2008). We note that theoretical studies of indirect similarity measures can also be found in the psychometric literature (e.g., Zegers and ten Berge, 1985). In this literature, the cosine is referred to as Tucker's congruence coefficient.

In the rest of this section, we focus our attention on the use of direct similarity measures. Direct similarity measures determine the similarity between two objects by taking the number of co-occurrences of the objects and adjusting this number for the total number of occurrences of each of the objects. In scientometric research, when a direct similarity measure is applied to co-occurrence data, the aim usually is to normalize the data, that is, to correct the data for differences in the number of occurrences of objects. This brings us to the main question of this paper: How should co-occurrence data be normalized? Or, in other words, which direct similarity measures are appropriate for normalizing co-occurrence data and which are not? We argue that co-occurrence data should always be normalized using a probabilistic similarity measure. Other direct similarity measures are not appropriate for normalization purposes. In particular,

set-theoretic similarity measures should not be used to normalize co-occurrence data.

To see why probabilistic similarity measures have the right properties for normalizing co-occurrence data, recall from Section 4 that the number of co-occurrences of two objects can be seen as the result of two independent effects, the similarity effect and the size effect. As we discussed in Section 4, probabilistic similarity measures correct for the size effect. This follows from Property 4.2. Set-theoretic similarity measures do not have this property, and they therefore do not properly correct for the size effect. As a consequence, set-theoretic similarity measures have, on average, higher values for objects that occur more frequently (see also Luukkonen et al., 1993; Zitt et al., 2000). The values of probabilistic similarity measures, on the other hand, do not depend on how frequently objects occur. This difference between set-theoretic and probabilistic similarity measures can easily be demonstrated empirically. In Figure 4, this is done for the term data set discussed in the previous section. (The author data set and the journal data set yield similar results.) The figure shows the relation between on the one hand the number of occurrences of a term and on the other hand the average similarity of a term with other terms. In the left panel of the figure, similarities are determined using a probabilistic similarity measure, namely the association strength. In this panel, there is no substantial correlation between the number of occurrences of a term and the average similarity of a term ($r = -0.069$, $\rho = -0.029$). This is very different in the right panel, in which similarities are determined using a set-theoretic similarity measure, namely the cosine. (The inclusion index and the Jaccard index yield similar results.) In the right panel, there is a strong positive correlation between the number of occurrences of a term and the average similarity of a term ($r = 0.839$, $\rho = 0.882$). Results such as those shown in the right panel clearly indicate that set-theoretic similarity measures do not properly correct for the size effect and, consequently, do not properly normalize co-occurrence data. It follows from this observation that one should be very careful with the interpretation of similarities that have been derived from co-occurrence data using a set-theoretic similarity measure (see also Luukkonen et al., 1993; Zitt et al., 2000). Moreover, when such similarities are analyzed using multivariate analysis techniques such as multidimensional scaling or hierarchical clustering, one should pay special attention to possible artifacts in the results of the analysis. When using multidimensional scaling, for example, it is our experience that frequently occurring objects tend to cluster together in the center of a solution.

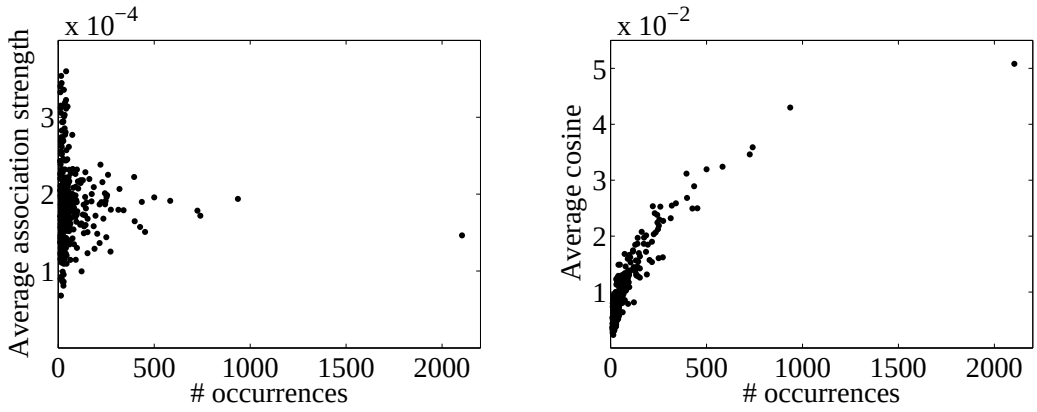

Figure 4: Relation between on the one hand the number of occurrences of a term and on the other hand the average similarity of a term with other terms. In the left panel, similarities are determined using the association strength. In the right panel, similarities are determined using the cosine.

To provide some additional insight why probabilistic similarity measures are more appropriate for normalization purposes than set-theoretic similarity measures, we now compare the main ideas underlying these two types of measures. Suppose that we are performing a co-word analysis and that we want to determine the similarity between two words, word i and word j. We consider two hypothetical scenarios, to which we refer as scenario 1 and scenario 2. The scenarios are summarized in Table 7, and they are illustrated graphically in the left and right panels of Figure 5. In each panel of the figure, the light gray rectangle represents the set of all documents used in the co-word analysis, the dark gray circle represents the set of all documents in which word i occurs, and the striped circle represents the set of all documents in which word j occurs. The area of a rectangle or circle is proportional to the number of documents in the corresponding set.

As can be seen in Table 7 and Figure 5, in scenario 1 words i and j both occur quite frequently, while in scenario 2 they both occur relatively infrequently. In both scenarios, however, the relative overlap of the set of documents in which word i occurs and the set of documents in which word j occurs is the same. That is, in both scenarios word i occurs in 30% of the documents in which word j occurs and, the other way around, word j occurs in 30% of the documents in which word i occurs. Because the relative overlap is the same, set-theoretic similarity measures, such as the cosine, the inclusion index, and the Jaccard index, yield the same similarity between words i and j in both scenarios (see Table 7). This is a consequence of Property 3.2

Table 7: Summary of two hypothetical scenarios in a co-word analysis.

	Scenario 1	Scenario 2
m	1 000	1 000
s_i	300	20
s_j	300	20
c_{ij}	90	6
Association strength	0.001	0.015
Cosine	0.300	0.300
Inclusion index	0.300	0.300
Jaccard index	0.176	0.176

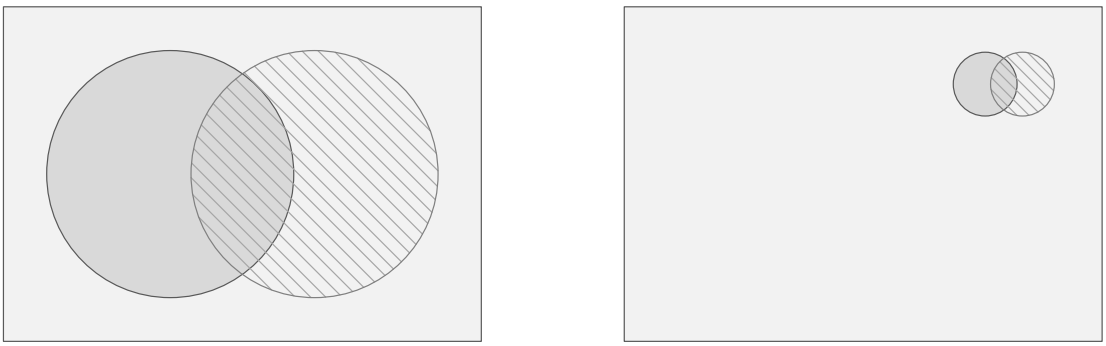

Figure 5: Graphical illustration of two hypothetical scenarios in a co-word analysis. Scenario 1 is shown in the left panel. Scenario 2 is shown in the right panel.

discussed in Section 3. At first sight, it might seem a natural result to have the same similarity between words i and j in both scenarios. However, we argue that this result is far from natural, at least for normalization purposes.

We first consider scenario 1 in more detail. In this scenario, words i and j each occur in 30% of all documents. If there is no special relation between words i and j and if, as a consequence, occurrences of the two words are statistically independent, one would expect the two words to co-occur in approximately $30\% \times 30\% = 9\%$ of all documents. As can be seen in Table 7, words i and j co-occur in exactly 9% of all documents. Hence, occurrences of words i and j seem to be statistically independent, at least approximately, and there seems to be no strong relation between the two words.

We now consider scenario 2. In this scenario, words i and j each occur in 2% of all documents. If occurrences of words i and j are statistically independent, one would expect the two words to co-occur in approximately 0.04% of all documents. However, words i and j co-occur in 0.6% of all documents, that is, they co-occur 15 times more frequently than would be expected under the assumption of statistical independence. Hence, there seems to be a quite strong relation between words i and j, definitely much stronger than in scenario 1.

It is clear that set-theoretic similarity measures yield results that do not properly reflect the difference between scenario 1 and scenario 2. This is because set-theoretic similarity measures are based on the idea of measuring the relative overlap of sets instead of the idea of measuring the deviation from statistical independence. Probabilistic similarity measures, such as the association strength, are based on the latter idea, and they therefore yield results that do properly reflect the difference between scenario 1 and scenario 2. As can be seen in Table 7, the association strength indicates that in scenario 2 the similarity between words i and j is 15 times higher than in scenario 1. This reflects that in scenario 2 the co-occurrence frequency of words i and j is 15 times higher than would be expected under the assumption of statistical independence while in scenario 1 the co-occurrence frequency of the two words equals the expected co-occurrence frequency under the independence assumption.

7 Conclusions

We have studied the application of direct similarity measures to co-occurrence data. Our survey of the scientometric literature has indicated that the most popular direct similarity measures are the association strength, the cosine, the inclusion index, and the Jaccard index. We have therefore focused most of our attention on these four measures. To make a well-considered decision which measure is most appropriate for ones purposes, we believe it to be indispensable to have a good theoretical understanding of the properties of the various measures. In this paper, we have analyzed these properties in considerable detail. Our analysis has revealed that there are two fundamentally different types of direct similarity measures. On the one hand, there are set-theoretic similarity measures, which can be interpreted as measures of the relative overlap of two sets. On the other hand, there are probabilistic similarity measures, which can be interpreted as measures of the deviation of observed co-occurrence frequencies from expected co-occurrence frequencies under an independence assumption. The cosine, the inclusion index, and the Jaccard index are examples of set-theoretic similarity measures, while the association strength is an example of a probabilistic similarity measure. Set-theoretic and probabilistic similarity measures serve different purposes, and it therefore makes no sense to argue that one measure is always better than another. In scientometric research, however, similarity measures are usually used for normalization purposes, and we have argued that in that specific case probabilistic similarity measures are much more appropriate than set-theoretic ones. Consequently, for most applications of direct similarity measures in scientometric research, we advise against the use of set-theoretic similarity measures and we recommend the use of a probabilistic similarity measure.

In addition to our theoretical analysis, we have also performed an empirical analysis of the behavior of various direct similarity measures. The analysis has shown that in practical applications the differences between various direct similarity measures can be quite large. This indicates that the issue of choosing an appropriate similarity measure is not only of theoretical interest but also has a high practical relevance. Another empirical observation that we have made is that set-theoretic similarity measures yield systematically higher values for frequently occurring objects than for objects that occur only a limited number of times. This confirms our theoretical finding that set-theoretic similarity measures do not properly correct for size effects. Probabilistic similarity measures do not have this problem.

There is one final comment that we would like to make. Above, we have argued in favor of the use of probabilistic similarity measures in scientometric research. Since probabilistic similarity measures are all proportional to each other, it does not really matter which probabilistic similarity measure one uses. In this paper, we have focused most of our attention on one particular probabilistic similarity measure, namely the association strength defined in (5). This measure shares with many other direct similarity measures the property that it takes values between zero and one. For practical purposes, however, it may be convenient not to use the measure in (5) directly but instead to multiply this measure by the number of documents m (e.g., Van Eck and Waltman, 2007; Van Eck et al., 2006). This results in a slight variant of the association strength. As discussed in Section 4, this variant has the appealing property that it equals one if the observed co-occurrence frequency of two objects equals the co-occurrence frequency that would be expected under the assumption that occurrences of the objects are statistically independent. It takes a value above or below one if the observed co-occurrence frequency is, respectively, higher or lower than the expected co-occurrence frequency under the independence assumption.

Appendix

In this appendix, we prove the theoretical results presented in the paper.

Proof of Proposition 3.1. We prove each property separately.

(Property 3.5) This property follows from Property 3.3. Property 3.3 implies that, if $c_{ij} > 0$, $S(c_{ij}, s_i, s_j) > S(c_{ij}, s_i + 1, s_j)$. Hence, if $c_{ij} > 0$, $S(c_{ij}, s_i, s_j)$ cannot take its minimum value. This means that $S(c_{ij}, s_i, s_j)$ can take its minimum value only if $c_{ij} = 0$. This proves Property 3.5.

(Property 3.6) This property follows from Properties 3.1, 3.2, and 3.3. Suppose that $c_{ij} = s_i = s_j$. For all $(c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$ such that $c'_{ij} = 0$, Property 3.1 implies that $S(c'_{ij}, s'_i, s'_j) \leq S(c_{ij}, s_i, s_j)$. For all $(c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$ such that $c'_{ij} > 0$, Property 3.3 implies that $S(c'_{ij}, s'_i, s'_j) \leq S(c'_{ij}, c'_{ij}, c'_{ij})$ and Property 3.2 implies that $S(c'_{ij}, c'_{ij}, c'_{ij}) = S(c_{ij}, s_i, s_j)$. Hence, for all $(c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$, $S(c'_{ij}, s'_i, s'_j) \leq S(c_{ij}, s_i, s_j)$. This means that, if $c_{ij} = s_i = s_j$, $S(c_{ij}, s_i, s_j)$ takes its maximum value. This proves Property 3.6.

(Property 3.7) This property follows from Properties 3.1, 3.3, and 3.5. Properties 3.1 and 3.5

imply that, if $c_{ij} = 0$, $S(c_{ij}, s_i, s_j)$ cannot take its maximum value. Property 3.3 implies that, if $0 < c_{ij} < s_i$ or $0 < c_{ij} < s_j$, $S(c_{ij}, s_i, s_j) < S(c_{ij}, c_{ij}, c_{ij})$. Hence, if $0 < c_{ij} < s_i$ or $0 < c_{ij} < s_j$, $S(c_{ij}, s_i, s_j)$ cannot take its maximum value. It now follows that $S(c_{ij}, s_i, s_j)$ can take its maximum value only if $c_{ij} = s_i = s_j$. This proves Property 3.7.

(Property 3.8) This property follows from Properties 3.1, 3.2, 3.3, and 3.5. If $c_{ij} = 0$, the property follows trivially from Properties 3.1 and 3.5. We therefore focus on the case in which $c_{ij} > 0$. Suppose, without loss of generality, that $0 < c_{ij} < s_i$. Consider an arbitrary constant $\alpha > 0$, and let $\beta = (c_{ij} + \alpha)/c_{ij}$. Property 3.2 implies that $S(\beta c_{ij}, \beta s_i, \beta s_j) = S(c_{ij}, s_i, s_j)$. Moreover, because $\beta c_{ij} = c_{ij} + \alpha$, $\beta s_i > s_i + \alpha$, and $\beta s_j \geq s_j + \alpha$, Property 3.3 implies that $S(\beta c_{ij}, \beta s_i, \beta s_j) < S(c_{ij} + \alpha, s_i + \alpha, s_j + \alpha)$. It now follows that $S(c_{ij} + \alpha, s_i + \alpha, s_j + \alpha) > S(c_{ij}, s_i, s_j)$. This proves Property 3.8. □

Proof of Proposition 3.2. Let $S(c_{ij}, s_i, s_j)$ denote an arbitrary set-theoretic similarity measure that has Property 3.9. We start by showing that for all $(c_{ij}, s_i, s_j), (c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$ the properties of set-theoretic similarity measures together with Property 3.9 are sufficient to determine whether $S(c_{ij}, s_i, s_j)$ is greater than, less than, or equal to $S(c'_{ij}, s'_i, s'_j)$. Suppose first that $c_{ij}, c'_{ij} > 0$. Let $\alpha = c_{ij}/c'_{ij}$. Property 3.2 implies that $S(\alpha c'_{ij}, \alpha s'_i, \alpha s'_j) = S(c'_{ij}, s'_i, s'_j)$. Moreover, taking into account that $c_{ij} = \alpha c'_{ij}$, it can be seen that Property 3.9 determines whether $S(c_{ij}, s_i, s_j)$ is greater than, less than, or equal to $S(\alpha c'_{ij}, \alpha s'_i, \alpha s'_j)$. Hence, if $c_{ij}, c'_{ij} > 0$, Properties 3.2 and 3.9 are sufficient to determine whether $S(c_{ij}, s_i, s_j)$ is greater than, less than, or equal to $S(c'_{ij}, s'_i, s'_j)$. Suppose next that $c_{ij} = 0$ or $c'_{ij} = 0$. Property 3.1 implies that $S(c_{ij}, s_i, s_j) = S(c'_{ij}, s'_i, s'_j)$ if $c_{ij} = c'_{ij} = 0$. Furthermore, Properties 3.1 and 3.5 imply that $S(c_{ij}, s_i, s_j) > S(c'_{ij}, s'_i, s'_j)$ if $c_{ij} > c'_{ij} = 0$ and, conversely, that $S(c_{ij}, s_i, s_j) < S(c'_{ij}, s'_i, s'_j)$ if $c'_{ij} > c_{ij} = 0$. Hence, if $c_{ij} = 0$ or $c'_{ij} = 0$, Properties 3.1 and 3.5 are sufficient to determine whether $S(c_{ij}, s_i, s_j)$ is greater than, less than, or equal to $S(c'_{ij}, s'_i, s'_j)$. It now follows that for all $(c_{ij}, s_i, s_j), (c'_{ij}, s'_i, s'_j) \in \mathcal{D}_S$ the properties of set-theoretic similarity measures together with Property 3.9 are sufficient to determine whether $S(c_{ij}, s_i, s_j)$ is greater than, less than, or equal to $S(c'_{ij}, s'_i, s'_j)$. This implies that all set-theoretic similarity measures that have Property 3.9 are monotonically related to each other. One of these measures is the cosine defined in (6). Hence, all set-theoretic similarity measures that have Property 3.9 are monotonically related to the cosine. This completes the proof of the proposition. □

Proof of Proposition 3.3. The proof is analogous to the proof of Proposition 3.2 provided above.

□

Proof of Proposition 3.4. Let $S(c_{ij}, s_i, s_j)$ denote an arbitrary weak set-theoretic similarity measure that has Property 3.11. Property 3.11 implies that, if $c_{ij} > 0$, $S(c_{ij}, s_i, s_j) > S(c_{ij}, s_i + 1, s_j + 1)$. Hence, if $c_{ij} > 0$, $S(c_{ij}, s_i, s_j)$ cannot take its minimum value. This means that $S(c_{ij}, s_i, s_j)$ can take its minimum value only if $c_{ij} = 0$. In other words, $S(c_{ij}, s_i, s_j)$ has Property 3.5. This shows that all weak set-theoretic similarity measures $S(c_{ij}, s_i, s_j)$ that have Property 3.11 also have Property 3.5. The rest of the proof is now analogous to the proof of Proposition 3.2 provided above. □

Proof of Proposition 3.5. It is easy to see that for all finite values of the parameter p the generalized similarity index defined in (10) has Properties 3.1, 3.2, and 3.3. Hence, it follows from Definition 3.1 that for all finite values of the parameter p the generalized similarity index is a set-theoretic similarity measure. This completes the proof of the proposition. □

Proof of Proposition 4.1. Let $S(c_{ij}, s_i, s_j)$ denote an arbitrary probabilistic similarity measure. Furthermore, let $c'_{ij} = c_{ij}/(s_i s_j)$ for all $i \neq j$, and let $s'_i = 1$ for all i. It follows from Property 4.2 that $S(c_{ij}, s_i, s_j) = S(c'_{ij}, s'_i, s'_j)$ for all $i \neq j$, and it follows from Property 4.1 that $S(c'_{ij}, s'_i, s'_j) = \alpha c'_{ij}$ for all $i \neq j$ and for some $\alpha > 0$. Hence, for all $i \neq j$ and for some $\alpha > 0$, $S(c_{ij}, s_i, s_j) = S(c'_{ij}, s'_i, s'_j) = \alpha c'_{ij} = \alpha c_{ij}/(s_i s_j) = \alpha S_{\mathrm{A}}(c_{ij}, s_i, s_j)$. In other words, $S(c_{ij}, s_i, s_j)$ is proportional to the association strength defined in (5). This completes the proof of the proposition. □

Proof of Corollary 4.2. The association strength defined in (5) does not have Property 3.2 and is therefore not a (weak or non-weak) set-theoretic similarity measure. The same is true for all measures that are proportional to the association strength. Consequently, it follows from Proposition 4.1 that a probabilistic similarity measure cannot also be a (weak or non-weak) set-theoretic similarity measure. This completes the proof of the corollary. □

References

P. Ahlgren, B. Jarneving, and R. Rousseau. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54(6):550–560, 2003.

M.R. Anderberg. *Cluster analysis for applications*. Academic Press, 1973.

F.B. Baulieu. A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, 6(1):233–246, 1989.

F.B. Baulieu. Two variant axiom systems for presence/absence based dissimilarity coefficients. *Journal of Classification*, 14(1):159–170, 1997.

I. Borg and P.J.F. Groenen. *Modern multidimensional scaling*. Springer, 2nd edition, 2005.

K.W. Boyack, R. Klavans, and K. Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005.

R.R. Braam, H.F. Moed, and A.F.J. van Raan. Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science*, 42(4):233–251, 1991a.

R.R. Braam, H.F. Moed, and A.F.J. van Raan. Mapping of science by combined co-citation and word analysis. II. Dynamical aspects. *Journal of the American Society for Information Science*, 42(4):252–266, 1991b.

M. Callon, J.P. Courtial, and F. Laville. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22(1):155–205, 1991.

Y.M. Chung and J.Y. Lee. A corpus-based approach to comparative evaluation of statistical term association measures. *Journal of the American Society for Information Science and Technology*, 52(4):283–296, 2001.

K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

M.A.A. Cox and T.F. Cox. Multidimensional scaling. In C. Chen, W. Härdle, and A. Unwin, editors, *Handbook of data visualization*, pages 315–347. Springer, 2008.

T.F. Cox and M.A.A. Cox. *Multidimensional scaling*. Chapman & Hall/CRC, 2nd edition, 2001.

D. de Solla Price. The analysis of scientometric matrices for policy implications. *Scientometrics*, 3(1):47–53, 1981.

F. Drasgow and L.E. Jones. Multidimensional scaling of derived dissimilarities. *Multivariate Behavioral Research*, 14(2):227–244, 1979.

L. Egghe. New relations between similarity measures for vectors based on vector norms. *Journal of the American Society for Information Science and Technology*, 2008. In press.

L. Egghe and C. Michel. Strong similarity measures for ordered sets of documents in information retrieval. *Information Processing and Management*, 38(6):823–848, 2002.

L. Egghe and C. Michel. Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques. *Information Processing and Management*, 39(5): 771–807, 2003.

L. Egghe and R. Rousseau. Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve. *Information Processing and Management*, 42(1):106–120, 2006.

W. Glänzel. National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1):69–115, 2001.

W. Glänzel, A. Schubert, and H.-J. Czerwon. A bibliometric analysis of international scientific cooperation of the European Union (1985–1995). *Scientometrics*, 45(2):185–202, 1999.

M. Gmür. Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57(1):27–57, 2003.

J.C. Gower. Measures of similarity, dissimilarity, and distance. In S. Kotz and N.L. Johnson, editors, *Encyclopedia of statistical sciences*, volume 5, pages 397–405. Wiley, 1985.

J.C. Gower and P. Legendre. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3(1):5–48, 1986.

J.P. Guilford. *Fundamental statistics in psychology and education*. McGraw-Hill, 5th edition, 1973.

L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Keters, R. Rousseau, and A. Vanhoutte. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management*, 25(3):315–318, 1989.

G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 2nd edition, 1952.

G. Heimeriks, M. Hörlesberger, and P. van den Besselaar. Mapping communication and collaboration in heterogeneous research networks. *Scientometrics*, 58(2):391–413, 2003.

S. Hinze. Bibliographical cartography of an emerging interdisciplinary discipline: The case of bioelectronics. *Scientometrics*, 29(3):353–376, 1994.

Z. Hubálek. Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews*, 57(4):669–689, 1982.

S. Janson and J. Vegelius. Measures of ecological association. *Oecologia*, 49(3):371–376, 1981.

B. Jarneving. A variation of the calculation of the first author cocitation strength in author cocitation analysis. *Scientometrics*, 2008. In press.

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

R. Klavans and K.W. Boyack. Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2):251–263, 2006a.

R. Klavans and K.W. Boyack. Quantitative evaluation of large maps of science. *Scientometrics*, 68(3):475–499, 2006b.

A. Kopcsa and E. Schiebel. Science and technology mapping: A new iteration model for representing multidimensional relationships. *Journal of the American Society for Information Science*, 49(1):7–17, 1998.

R.N. Kostoff, H.J. Eberhart, and D.R. Toothman. Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography. *Journal of the American Society for Information Science*, 50(5):427–447, 1999.

R.N. Kostoff, J.A. del Río, J.A. Humenik, E.O. García, and A.M. Ramírez. Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*, 52(13):1148–1156, 2001.

J. Law and J. Whittaker. Mapping acidification research: A test of the co-word method. *Scientometrics*, 23(3):417–461, 1992.

M. Leclerc and J. Gagné. International scientific cooperation: The continentalization of science. *Scientometrics*, 31(3):261–292, 1994.

L. Leydesdorff. Words and co-words as indicators of intellectual organization. *Research Policy*, 18(4):209–223, 1989.

L. Leydesdorff. On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85, 2008.

L. Leydesdorff and L. Vaughan. Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. *Journal of the American Society for Information Science and Technology*, 57(12):1616–1628, 2006.

T. Luukkonen, O. Persson, and G. Sivertsen. Understanding patterns of international scientific collaboration. *Science, Technology, and Human Values*, 17(1):101–126, 1992.

T. Luukkonen, R.J.W. Tijssen, O. Persson, and G. Sivertsen. The measurement of international scientific collaboration. *Scientometrics*, 28(1):15–36, 1993.

C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

K.W. McCain. Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6):433–443, 1990.

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

K.W. McCain. The structure of biotechnology R & D. *Scientometrics*, 32(2):153–175, 1995.

F. Morillo, M. Bordons, and I. Gómez. 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, 2003.

C.L. Palmer. Structures and strategies of interdisciplinary science. *Journal of the American Society for Information Science*, 50(3):242–253, 1999.

H.P.F. Peters and A.F.J. van Raan. Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. *Research Policy*, 22(1):23–45, 1993a.

H.P.F. Peters and A.F.J. van Raan. Co-word-based science maps of chemical engineering. Part II: Representations by combined clustering and multidimensional scaling. *Research Policy*, 22(1):23–45, 1993b.

H.P.F. Peters, R.R. Braam, and A.F.J. van Raan. Cognitive resemblance and citation relations in chemical engineering publications. *Journal of the American Society for Information Science*, 46(1):9–21, 1995.

J. Qin. Semantic similarities between a keyword database and a controlled vocabulary database: An investigation in the antibiotic resistance literature. *Journal of the American Society for Information Science*, 51(3):166–180, 2000.

A. Rip and J.-P. Courtial. Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, 6(6):381–400, 1984.

M. Rorvig. Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science*, 50(8):639–651, 1999.

S. Rosenberg and R. Jones. A method for investigating and representing a person's implicit theory of personality: Theodore Dreiser's view of people. *Journal of Personality and Social Psychology*, 22(3):372–386, 1972.

S. Rosenberg, C. Nelson, and P.S. Vivekananthan. A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, 9(4):283–294, 1968.

G. Salton. Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, 10(4):440–457, 1963.

G. Salton and M.J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, 1983.

J.W. Schneider and P. Borlund. Matrix comparison, part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology*, 58(11):1586–1595, 2007a.

J.W. Schneider and P. Borlund. Matrix comparison, part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics. *Journal of the American Society for Information Science and Technology*, 58(11):1596–1609, 2007b.

J.W. Schneider, B. Larsen, and P. Ingwersen. A comparative study of first and all-author co-citation counting, and two different matrix generation approaches applied for author co-citation analyses. *Scientometrics*, 2008. In press.

A. Schubert and T. Braun. International collaboration in the sciences, 1981–1985. *Scientometrics*, 19(1–2):3–10, 1990.

M.W. Simmen. Multidimensional scaling of binary dissimilarities: Direct and derived approaches. *Multivariate Behavioral Research*, 31(1):47–67, 1996.

H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.

H. Small. The relationship of information science to the social sciences: A co-citation analysis. *Information Processing and Management*, 17(1):39–50, 1981.

H. Small. A SCI-Map case study: Building a map of AIDS research. *Scientometrics*, 30(1): 229–241, 1994.

H. Small and E. Greenlee. Citation context analysis of a co-citation cluster: Recombinant-DNA. *Scientometrics*, 2(4):277–301, 1980.

H. Small and E. Sweeney. Clustering the science citation index using co-citations. I. A comparison of methods. *Scientometrics*, 7(3–6):391–409, 1985.

H. Small, E. Sweeney, and E. Greenlee. Clustering the science citation index using co-citations. II. Mapping science. *Scientometrics*, 8(5–6):321–340, 1985.

R.R. Sokal and P.H.A. Sneath. *Principles of numerical taxonomy*. Freeman, 1963.

R.J.W. Tijssen. A quantitative assessment of interdisciplinary structures in science and technology: Co-classification analysis of energy research. *Research Policy*, 21(1):27–44, 1992.

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

R.J.W. Tijssen and A.F.J. van Raan. Mapping co-word structures: A comparison of multidimensional scaling and LEXIMAPPE. *Scientometrics*, 15(3–4):283–295, 1989.

W.A. van der Kloot and H. van Herk. Multidimensional scaling of sorting data: A comparison of three procedures. *Multivariate Behavioral Research*, 26(4):563–581, 1991.

N.J. van Eck and L. Waltman. Bibliometric mapping of the computational intelligence field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 15(5):625–645, 2007.

N.J. van Eck and L. Waltman. Appropriate similarity measures for author co-citation analysis. *Journal of the American Society for Information Science and Technology*, 59(10):1653–1661, 2008.

N.J. van Eck, L. Waltman, J. van den Berg, and U. Kaymak. Visualizing the computational intelligence field. *IEEE Computational Intelligence Magazine*, 1(4):6–10, 2006.

A.F.J. van Raan and R.J.W. Tijssen. The neural net of neural network research: An exercise in bibliometric mapping. *Scientometrics*, 26(1):169–192, 1993.

L. Vaughan. Visualizing linguistic and cultural differences using Web co-link data. *Journal of the American Society for Information Science and Technology*, 57(9):1178–1193, 2006.

L. Vaughan and J. You. Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. *Scientometrics*, 68(3):611–628, 2006.

L. Waltman and N.J. van Eck. Some comments on the question whether co-occurrence data should be normalized. *Journal of the American Society for Information Science and Technology*, 58(11):1701–1703, 2007.

M.J. Warrens. *Similarity coefficients for binary data*. PhD thesis, Leiden University, 2008.

H.D. White and B.C. Griffith. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3):163–171, 1981.

H.D. White and K.W. McCain. Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, 49(4):327–355, 1998.

F.E. Zegers and J.M.F. ten Berge. A family of association coefficients for metric scales. *Psychometrika*, 50(1):17–24, 1985.

M. Zitt, E. Bassecoulard, and Y. Okubo. Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics*, 47(3):627–657, 2000.

Publications in the Report Series Research* in Management

ERIM Research Program: "Business Processes, Logistics and Information Systems"

2009

How to Normalize Co-Occurrence Data? An Analysis of Some Well-Known Similarity Measures
Nees Jan van Eck and Ludo Waltman
ERS-2009-001-LIS
http://hdl.handle.net/1765/14528

---

* A complete overview of the ERIM Report Series Research in Management:
https://ep.eur.nl/handle/1765/1

ERIM Research Programs:
LIS Business Processes, Logistics and Information Systems
ORG Organizing for Performance
MKT Marketing
F&A Finance and Accounting
STR Strategy and Entrepreneurship