Made available by Hasselt University Library in https://documentserver.uhasselt.be

Good Properties of Similarity Measures and Their Complementarity

Peer-reviewed author version

EGGHE, Leo (2010) Good Properties of Similarity Measures and Their Complementarity. In: JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 61(10). p. 2151-2160.

DOI: 10.1002/asi.21380
Handle: http://hdl.handle.net/1942/11317

Good properties of similarity measures and their complementarity

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium
leo.egghe@uhasselt.be

ABSTRACT

Similarity measures, such as the ones of Jaccard, Dice or Cosine, measure the similarity between two vectors. A good property for similarity measures would be that, if we add a constant vector to both vectors, then the similarity must increase. We show that Dice and Jaccard satisfy this property while Cosine and both overlap measures do not. Adding a constant vector is called, in Lorenz concentration theory, "nominal increase" and we show that the stronger "transfer principle" is not a required good property for similarity measures.

Another good property is that, when we have two vectors and if we add one of these vectors to both vectors, then the similarity must increase. Now Dice, Jaccard, Cosine and one of the overlap measures satisfy this property, while the other overlap measure does not. Also a variant of this latter property is studied.

---

[1] Permanent address

Acknowledgement : The author is grateful to L. Waltman for the example in the second section, the proof in the third section and for interesting discussions on the topic of the paper.
Key words and phrases: nominal increase, Jaccard, Dice, Cosine, overlap, Lorenz.

I. Introduction

Similarity measures are applied on cooccurrence data or cocitation data of authors and journals, represented by vectors of the form $\vec{X} = (x_1, x_2, ..., x_N)$ and $\vec{Y} = (y_1, y_2, ..., y_N)$. Some papers (e.g. van Eck and Waltman (2009)) only consider binary vectors where the coordinates of $\vec{X}$ and $\vec{Y}$ are 0 or 1. This limitation will not be applied here but we will, evidently, assume that $x_i, y_i \geq 0$ for all $i = 1, ..., N$.

Let us repeat the "classical" similarity measures. First there is Dice's measure E

$$E = \frac{2\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} y_i^2} \tag{1}$$

Since all sums are $\sum_{i=1}^{N}$, we will, henceforth, use the simpler $\sum$. Jaccard's measure, denoted J is

$$J = \frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i y_i} \tag{2}$$

Both measures are, essentially, the same since it is not difficult to prove that

$$E = \frac{2J}{J+1} \tag{3}$$

so that E is an increasing function of J (and vice-versa). Hence all properties expressed with inequalities for one measure are also valid for the other measure.

The cosine formula (e.g. used by Salton in information retrieval – Salton and McGill (1987)) is defined as follows.

$$C = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \tag{4}$$

One can, indeed, show that C is the cosine of the angle between the vectors $\vec{X}$ and $\vec{Y}$.

The classical overlap measures O_1 and O_2 are defined as follows.

$$O_1 = \frac{\sum x_i y_i}{\min\left(\sum x_i^2, \sum y_i^2\right)} \tag{5}$$

$$O_2 = \frac{\sum x_i y_i}{\max\left(\sum x_i^2, \sum y_i^2\right)} \tag{6}$$

The name overlap comes from the binary case where $\sum x_i y_i$ can be interpreted as the number of elements in the intersection of two sets, each set having $\sum x_i^2$ (resp. $\sum y_i^2$) elements (see Egghe (2009) for a complete description). In an information retrieval situation, (5) and (6) (in this or another order) can be interpreted as the classical recall and precision measures.

Similarity measures are also studied (or mentioned) in the books Boyce, Meadow and Kraft (1995), Grossman and Frieder (1998), Losee (1998), Salton and McGill (1987), Tague-Sutcliffe (1995) and van Rijsbergen (1979); also see Egghe and Michel (2002, 2003). In van Eck and Waltman (2009) a general similarity measure with a parameter p is defined such that it yields all measures (except J) defined above according to some (limiting) values of p (see the last section for more details).

Strictly speaking the correlation coefficient of Pearson, defined in (7), as also a similarity measure

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}\sqrt{\sum\left(y_i - \bar{y}\right)^2}} \tag{7}$$

where $\bar{x}$ and $\bar{y}$ denote the average of $\vec{X}$ and $\vec{Y}$, respectively.

Certainly, r is an important measure of correlation in statistics (measuring e.g. the degree of correlation of a cloud of points with its regression line). But in Ahlgren, Jarneving and Rousseau (2003) one has found two bad properties of r in terms of similarity. One of them is that, adding the same number of zeros to two vectors should not decrease $(\geq)$ the similarity. It is shown in Ahlgren, Jarneving and Rousseau (2003) that r does not satisfy this property. The second property is defined in a similar way and also here we have that r does not satisfy this property. So we will not consider r anymore in this study of similarity measures (note that all the measures defined above (except r) are invariant to adding zeros to vectors).

The idea in Ahlgren, Jarneving and Rousseau (2003) of checking good properties for similarity measures is comparable with (but not the same as) what has been done in econometrics and informetrics on so-called concentration measures. There are, however, basic differences. Firstly, concentration measures act on <u>one</u> vector $\vec{X}$ and where one tries to measure the degree of inequality amongst the coordinates of $\vec{X}$: the higher the inequality, the higher the value of the concentration measures. We refer to Egghe (2005) and references therein for many applications of concentration theory in informetrics. Similarity measures, however, compare <u>two</u> vectors which makes them very different from concentration measures. Secondly, similarity measures give high values for vectors $\vec{X}$ and $\vec{Y}$ that are similar (e.g. $\vec{X} = \vec{Y} \Rightarrow$ highest value for the similarity measure) while concentration measures give high values for a vector $\vec{X}$ with concentrated coordinate values (e.g. $\vec{X} = (1,0,...,0)$). This last remark is not completely clear, stated as such, since concentration measures act on one vector and similarity measures act on two vectors, but let us clarify this further.

Let f be a concentration measure. Let $\vec{X}$ be any vector (say with positive coordinates, including zero). Let $\vec{A} = (a,...,a)$ be a constant vector with $a > 0$. Then it is logical that the vector $\vec{X} + \vec{A} = (x_1 + a, x_2 + a, ..., x_N + a)$ is less concentrated than $\vec{X}$ (supposing $\vec{X}$ not to be a constant vector), i.e. $f\left(\vec{X} + \vec{A}\right) < f\left(\vec{X}\right)$. In concentration theory this is called the "principle of nominal increase". Let F be a similarity measure, hence acting on two vectors $\left(\vec{X}, \vec{Y}\right)$. It is

then logical that (supposing $\vec{X} \neq \vec{Y}$) that $\vec{X} + \vec{A}$ and $\vec{Y} + \vec{A}$ are more similar than $\vec{X}$ and $\vec{Y}$: $F\left(\vec{X} + \vec{A}, \vec{Y} + \vec{A}\right) > F\left(\vec{X}, \vec{Y}\right)$. So similarity measures are kind of "opposite" to concentration measures but acting on two vectors instead of one for concentration measures. This is only partially true as we will explain in the next section.

Also in the next section we will study this principle of nominal increase for similarity measures. We will show that Dice's measure as well as the Jaccard index satisfy this principle, while Cosine, and overlap measure O_1 and O_2 do not satisfy this principle. For O_2 we have partial positive results if $\vec{X}$ is a multiple of $\vec{Y}$ or if $\sum x_i = \sum y_i$.

In the third section we will study another "natural" good property for similarity measures. If we add to two vectors $\vec{X}$ and $\vec{Y}$ one of the two vectors, say $\vec{X} + \vec{Y}$ and $\vec{Y} + \vec{Y} = 2\vec{Y}$ then the similarity should increase strictly (if $\vec{X} \neq \vec{Y}$). We will show that Dice, Jaccard, Cosine and overlap measure O_2 satisfy this property while overlap measure O_1 does not satisfy this property. A variant of the property studied in this section is also defined and studied and we prove that only Cosine satisfies this property.

We hereby demonstrated a kind of complementarity of the classical similarity measures : some measures satisfy a property while others do not, while the reverse is valid for another property !

The paper then closes with suggestions for further research. We can generalise both properties in Section 2 and 3 by requiring that $\vec{X} + \vec{Z}$ and $\vec{Y} + \vec{Z}$ are more similar than $\vec{X}$ and $\vec{Y}$ themselves (if $\vec{X} \neq \vec{Y}$) for every vector $\vec{Z}$ (with positive coordinates). Dice and Jaccard satisfy this property while the other measures do not.

II. The principle of nominal increase for similarity measures

Let sim be any measure we want to check for good properties of similarity measures. Let $\vec{X} = (x_1,...,x_N)$, $\vec{Y} = (y_1,...,y_N)$ with $x_i, y_i \geq 0$ for all $i = 1,...,N$ and such that $\vec{X} \neq \vec{Y}$. Let $\vec{A} = (a,...,a)$ be any constant vector with $a > 0$. Then we say that sim satisfies the principle of nominal increase if

$$sim\left(\vec{X} + \vec{A}, \vec{Y} + \vec{A}\right) > sim\left(\vec{X}, \vec{Y}\right) \tag{8}$$

The following theorem proves that Dice's measure satisfies this principle.

Theorem II.1: Dice's measure E (formula (1)) satisfies the principle of nominal increase.

Proof: Denote

$$E(a) = \frac{2\sum (x_i + a)(y_i + a)}{\sum (x_i + a)^2 + \sum (y_i + a)^2} \tag{9}$$

being Dice's measure on $\left(\vec{X} + \vec{A}, \vec{Y} + \vec{A}\right)$. For proving that $E(a) > E$ it is sufficient to prove that $\frac{dE(a)}{da} > 0$ (this method will not always be successful since the principle of nominal increase can be valid without having a strictly increasing function of a). We have

$$\frac{dE(a)}{da} = \frac{1}{\left(\sum (x_i + a)^2 + \sum (y_i + a)^2\right)^2}[*]$$

where

$$* = \left(\sum (x_i + a)^2 + \sum (y_i + a)^2\right)\left(2\sum x_i + 2\sum y_i + 4Na\right)$$
$$-2\sum (x_i + a)(y_i + a)\left(2\sum (x_i + a) + 2\sum (y_i + a)\right)$$

So $\frac{dE(a)}{da} > 0$ if and only if

$$\left(\sum (x_i + a)^2 + \sum (y_i + a)^2\right)\left(2\sum (x_i + a) + 2\sum (y_i + a)\right)$$
$$-4\sum (x_i + a)(y_i + a)\left(\sum (x_i + a) + \sum (y_i + a)\right) > 0$$

if and only if

$$2\left(\sum (x_i + a)^2 - 2\sum (x_i + a)(y_i + a) + \sum (y_i + a)^2\right)\left(\sum (x_i + a) + \sum (y_i + a)\right) > 0$$

if and only if

$$\left(\sum \left((x_i + a) - (y_i + a)\right)^2\right)\left(\sum (x_i + a) + \sum (y_i + a)\right) > 0$$

which is true since all coordinates are ≥ 0, $a > 0$ and $\vec{X} \neq \vec{Y}$. □

Also Jaccard's index J satisfies this principle.

Theorem II.2: Jaccard's measure J (formula (2)) satisfies the principle of nominal increase.

First proof: Denote

$$J(a) = \frac{\sum (x_i + a)(y_i + a)}{\sum (x_i + a)^2 + \sum (y_i + a)^2 - \sum (x_i + a)(y_i + a)} \qquad (10)$$

being Jaccard's index on $\left(\vec{X}+\vec{A},\vec{Y}+\vec{A}\right)$. The same method as in the proof of Theorem II.1 yields $\frac{dJ(a)}{da}>0$ if and only if

$$\left(\sum\left((x_i+a)-(y_i+a)\right)^2\right)\left(\sum(x_i+a)+\sum(y_i+a)\right)>0$$

which is true.

Second proof: By Theorem II.1 we have $E(a)>E$. It follows from (3) that

$$J=\frac{E}{2-E} \tag{11}$$

a strictly increasing function of E. Hence

$$J(a)=\frac{E(a)}{2-E(a)}>\frac{E}{2-E}=J\,. \qquad \square$$

Now the Cosine measure does not satisfy the principle of nominal increase.

Example II.3: Let $\vec{X}=(3,1)$, $\vec{Y}=(6,2)$ (hence $\vec{Y}=2\vec{X}$), $\vec{A}=(1,1)$. Then the similarity between $\vec{X}$ and $\vec{Y}$ is 1 of course. But on $\left(\vec{X}+\vec{A},\vec{Y}+\vec{A}\right)$ we have

$$C(a)=\frac{34}{\sqrt{20}\sqrt{58}}=0.9982744<1.$$

It is easy to see that

$$O_1O_2=C^2 \tag{12}$$

This already yields that at least one of the measures O_1 or O_2 do not satisfy the principle of nominal increase. The same example yields $O_1 = 2$ and $O_1(a) = \frac{34}{20} < 2$, a counterexample.

For O_2 this is not a counterexample: $O_2 = \frac{1}{2} < O_2(a) = \frac{34}{58}$, but also for O_2 a counterexample exists.

This counterexample was kindly provided by L. Waltman (based on simulations in Mathlab).

$\vec{X} = (9636, 6142, 7457, 5318, 59)$,

$\vec{Y} = (9246, 7469, 6486, 5098, 2188)$

$\vec{A} = (a, a, a, a, a)$ with $a = 4867$.

Then it is readily seen that $O_2 = 0.9818$ (between $\vec{X}$ and $\vec{Y}$) while

$O_2(a) = 0.9799 < O_2(O_2(a))$ is the O_2-measure for $(\vec{X} + \vec{A}, \vec{Y} + \vec{A})$

As a partial position result we can show that O_2 satisfies this principle if one vector is a multiple of the other vector or in case the sum of the coordinates of the vectors are equal. We will prove this now.

Proposition II.4: Let $\vec{X}$, $\vec{Y}$ and $\vec{A}$ be as above. Suppose $\vec{Y} = \alpha \vec{X}$ for a certain number $\alpha > 0$ $\alpha \neq 1$ (excluding the case $\vec{Y} = \vec{X}$ in which case the principle of nominal increase is never valid !). Then

$$O_2\left(\vec{X} + \vec{A}, \vec{Y} + \vec{A}\right) > O_2\left(\vec{X}, \vec{Y}\right) = O_2$$

Proof: It is no loss of generality to suppose $\alpha > 1$ (otherwise, interchange $\vec{X}$ and $\vec{Y}$). Then it is easy to see that

$$O_2 = \frac{1}{\alpha} \tag{13}$$

and that

$$O_2\left(\vec{X}+\vec{A},\vec{Y}+\vec{A}\right)=\frac{\sum\left(x_i+a\right)\left(\alpha x_i+a\right)}{\sum\left(\alpha x_i+a\right)^2}$$

since all $x_i \geq 0$. Now

$$\frac{\sum\left(x_i+a\right)\left(\alpha x_i+a\right)}{\sum\left(\alpha x_i+a\right)^2}>\frac{1}{\alpha}$$

if and only if

$$a\sum x_i+\alpha a\sum x_i+Na^2>2a\sum x_i+\frac{1}{\alpha}Na^2$$

if and only if

$$\sum x_i\left(\alpha a-a\right)>\left(\frac{1}{\alpha}-1\right)Na^2$$

which is true since $\alpha>1$. □

Note that the above Proposition is false for O_1 (see the example above). In fact, in all cases where $\vec{Y}=\alpha\vec{X}$ $\left(\alpha>0,\alpha\neq 1\right)$ we have that $C\left(\vec{X},\vec{Y}\right)=1$. Then by (12) and the above Proposition on O_2 we have that $O_1\left(\vec{X}+\vec{A},\vec{Y}+\vec{A}\right)<O_1\left(\vec{X},\vec{Y}\right)$. But also note that this inequality is not always true for O_1. Indeed, take the example $\vec{X}=\left(1.1\frac{\sqrt{2}}{2};1.1\frac{\sqrt{2}}{2}\right)$, $\vec{Y}=\left(0;0.9\sqrt{2}\right)$ and $\vec{A}=\left(1,1\right)$. Then $O_1\left(\vec{X},\vec{Y}\right)=0.8181...<O_1\left(\vec{X}+\vec{A},\vec{Y}+\vec{A}\right)=0.9436944$.

In order to prove a partial result on the principle of nominal increase for O_2 we need to show

$$\frac{\sum\left(x_i+a\right)\left(y_i+a\right)}{\max\left(\sum\left(x_i+a\right)^2,\sum\left(y_i+a\right)^2\right)}>\frac{\sum x_i y_i}{\max\left(\sum x_i^2,\sum y_i^2\right)} \qquad (14)$$

We can always suppose that $\sum x_i^2 \le \sum y_i^2$ (otherwise interchange $\vec{X}$ and $\vec{Y}$). However this does not yield that $\sum (x_i + a)^2 \le \sum (y_i + a)^2$ for all $a > 0$. To produce a counterexample to this we must find $\vec{X}$, $\vec{Y}$ and $\vec{A}$ such that

$$\sum x_i^2 + 2a\sum x_i + a^2 > \sum y_i^2 + 2a\sum y_i + a^2$$

and hence, since $\sum x_i^2 \le \sum y_i^2$, we must have $\sum x_i > \sum y_i$. In terms of L^1- and L^2-vector norms we must have $\| \vec{X} \|_2 \le \| \vec{Y} \|_2$ and $\| \vec{X} \|_1 > \| \vec{Y} \|_1$. It is well-known that equal $\| . \|_2$-norms yield a circle and that equal $\| . \|_1$-norms yield a square. Let us take the circle $\| \vec{Z} \|_2 = 1.2$ and the square $\| \vec{Z} \|_1 = \sqrt{2}$. Then we have that they intersect (see Fig. 1).

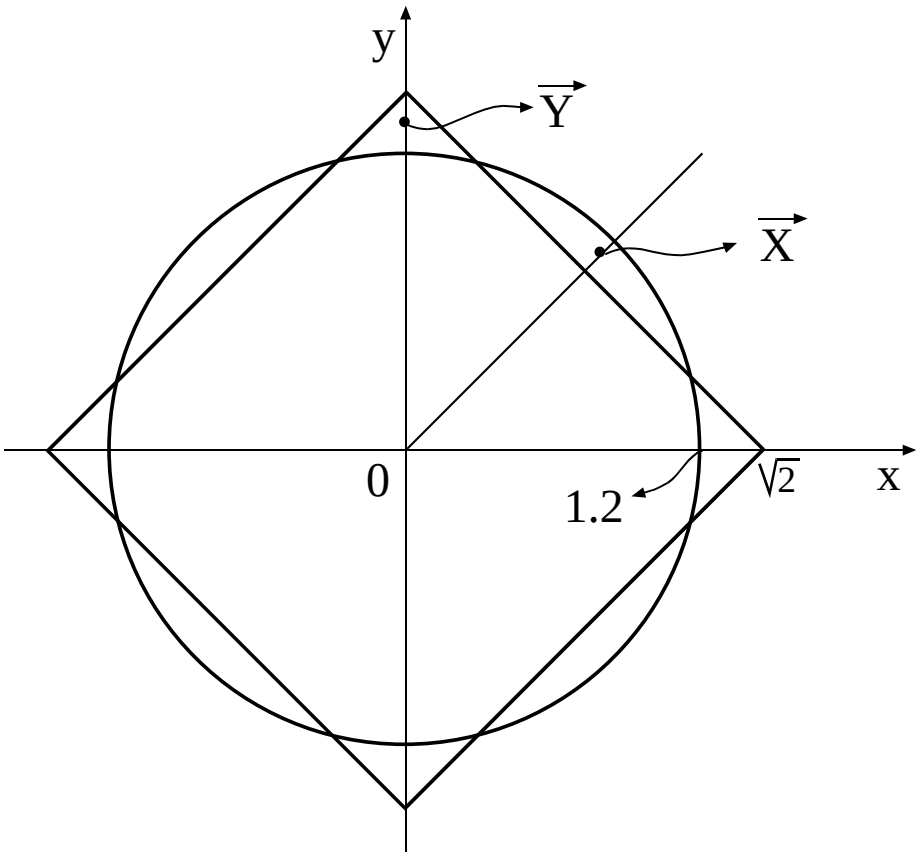

Fig. 1. Circle $x^2 + y^2 = (1.2)^2$ and square $x + y = \sqrt{2}$ yielding $\vec{X}$, $\vec{Y}$ such that $\| \vec{X} \|_2 < \| \vec{Y} \|_2$ and $\| \vec{X} \|_1 > \| \vec{Y} \|_1$.

Take $\vec{X} = \left(1.1\frac{\sqrt{2}}{2}; 1.1\frac{\sqrt{2}}{2}\right)$ and $\vec{Y} = \left(0; 0.9\sqrt{2}\right)$. Then we have that

$\sum x_i^2 = 1.21 < \sum y_i^2 = 1.62$ and $\sum x_i = 1.1\sqrt{2} > \sum y_i = 0.9\sqrt{2}$, i.e. $\vec{X}$ is inside the circle $\| \vec{Z} \|_2 = 1.2$ and outside the square $\| \vec{Z} \|_1 = \sqrt{2}$ while we have the opposite for $\vec{Y}$.

(Incidentally we could use this example above in the O_1 study but the example was made for the purpose we have here: study further the principle of nominal increase for O_2).

If we take $\vec{A} = (1,1)$ then, with $\vec{X}$ and $\vec{Y}$ as above, we have

$$\sum (x_i + a)^2 = 5.8284271 < \sum (y_i + a)^2 = 6.1655844$$

but with $\vec{A} = (2,2)$ we have

$$\sum (x_i + a)^2 = 15.43254 > \sum (y_i + a)^2 = 11.71627 .$$

We can now state and prove the following Proposition.

Proposition II.5: Let $\vec{X}$, $\vec{Y}$ and $\vec{A}$ be as above. Suppose $\vec{X} \neq \vec{Y}$ and

$$\| \vec{X} \|_1 = \sum x_i = \sum y_i = \| \vec{Y} \|_1 \qquad (15)$$

Suppose $\vec{X}$ is not a multiple of $\vec{Y}$ (as in Proposition II.4). Then

$$O_2\left(\vec{X} + \vec{A}, \vec{Y} + \vec{A}\right) > O_2\left(\vec{X}, \vec{Y}\right) = O_2$$

Proof: We can always suppose $\sum x_i^2 \le \sum y_i^2$ (otherwise interchange $\vec{X}$ and $\vec{Y}$). Since we have to prove (14) we hence must show

$$\frac{\sum (x_i + a)(y_i + a)}{\max\left(\sum (x_i + a)^2, \sum (y_i + a)^2\right)} > \frac{\sum x_i y_i}{\sum y_i^2}$$

In view of the $\|.\|_1$ and $\|.\|_2$ argument above, this is equivalent with

$$\frac{\sum (x_i + a)(y_i + a)}{\sum (y_i + a)^2} > \frac{\sum x_i y_i}{\sum y_i^2} \tag{16}$$

and

$$\frac{\sum (x_i + a)(y_i + a)}{\sum (x_i + a)^2} > \frac{\sum x_i y_i}{\sum y_i^2} \tag{17}$$

For (16) we must have

$$\left(\sum x_i y_i + a\sum x_i + a\sum y_i + Na^2\right)\left(\sum y_i^2\right) > \left(\sum y_i^2 + 2a\sum y_i + Na^2\right)\left(\sum x_i y_i\right)$$

or

$$\left(\sum x_i\right)\left(\sum y_i^2\right) + \left(\sum y_i\right)\left(\sum y_i^2\right) + Na\left(\sum y_i^2\right) > 2\left(\sum y_i\right)\left(\sum x_i y_i\right) + Na\left(\sum x_i y_i\right) \tag{18}$$

Since $\vec{X}$ is not a multiple of $\vec{Y}$ we have that

$$\sum x_i y_i < \sqrt{\sum x_i^2}\sqrt{\sum y_i^2} \le \sum y_i^2 \tag{19}$$

So, if

$$\sum y_i \le \sum x_i \tag{20}$$

we have that (18) (with strict inequality) is valid.

For (17) we must have

$$\left(\sum x_i y_i + a\sum x_i + a\sum y_i + Na^2\right)\left(\sum y_i^2\right) > \left(\sum x_i^2 + 2a\sum x_i + Na^2\right)\left(\sum x_i y_i\right)$$

or

$$\left(\sum x_i y_i\right)\left(\sum y_i^2\right) + a\left(\sum x_i\right)\left(\sum y_i^2\right) + a\left(\sum y_i\right)\left(\sum y_i^2\right) + Na^2\left(\sum y_i^2\right)$$

$$> \left(\sum x_i y_i\right)\left(\sum x_i^2\right) + 2a\left(\sum x_i y_i\right)\left(\sum x_i\right) + Na^2\left(\sum x_i y_i\right) \tag{21}$$

But since $\sum x_i^2 \le \sum y_i^2$ and by (19), the inequality (21) is proved if

$$\sum x_i \le \sum y_i \tag{22}$$

(20) and (22) show that the Proposition is proved if we have (15). □

A variant of this proof is as follows. We can suppose that $\sum x_i^2 \le \sum y_i^2$. For $\sum (x_i + a)^2 \le \sum (y_i + a)^2$ we need

$$\sum x_i^2 + 2a\sum x_i + a^2 \le \sum y_i^2 + 2a\sum y_i + a^2$$

If we suppose that $\sum x_i \le \sum y_i$ then this is satisfied. By (15) we now only have to prove (16). We take over from the first proof the calculation, showing that (16) is valid if $\sum y_i \le \sum x_i$. Hence Proposition II.5 is proved, supposing $\sum x_i = \sum y_i$.

Remark: We have claimed that the principle of nominal increase is a good property for similarity measures. This principle, however is, in concentration theory, linked with the transfer principle as follows. Let f be a concentration measure. As briefly described in the Introduction, it satisfies the principle of nominal increase (acting on one vector): $f\left(\vec{X}+\vec{A}\right)<f\left(\vec{X}\right)$. The transfer principle is defined as follows (see e.g. Egghe and Rousseau (1990), Egghe (2005)): If, in $\vec{X}$ we have two coordinates such that $x_j \le x_i$ $\left(i,j=1,...,N\right)$ and if $\vec{X}'$ is this vector derived from $\vec{X}$ such that all coordinates are the same as in $\vec{X}$ except the ith and the jth one: x_j becomes x_j-h (with $h>0$ such that $x_j-h\ge 0$) and x_i becomes x_i+h, then $\vec{X}'$ is more concentrated than $\vec{X}$ and we must have $f\left(\vec{X}'\right)>f\left(\vec{X}\right)$.

In econometric terms: the "poorer" person j becomes poorer and the "richer" person i becomes richer. This operation is also called an elementary transfer.

One can prove that the transfer principle implies the principle of nominal increase for concentration measures. One can now wonder if the transfer principle, applied to similarity measures, is a good property. In this context, this would mean that, given two vectors $\vec{X}$, $\vec{Y}$ and suppose that $\vec{X}'$ and $\vec{Y}'$ are constructed from $\vec{X}$, respectively $\vec{Y}$ by an elementary transfer on the same coordinates (supposing $x_j \le x_i$ and $y_j \le y_i$) the similarity has diminished then: let F be such a similarity measure, should we then have $F\left(\vec{X}',\vec{Y}'\right)<F\left(\vec{X},\vec{Y}\right)$?

The answer is no and this is logical as the next example shows. Let $\vec{X}=(3,2)$, $\vec{Y}=(4,2)$, $\vec{X}'=(4,1)$, $\vec{Y}'=(5,1)$. So both vectors $\vec{X}'$, $\vec{Y}'$ are derived from $\vec{X}$ and $\vec{Y}$ by an elementary transfer (of one unit) from the second coordinate to the first coordinate. However, it is easy to verify that

$$C\left(\vec{X},\vec{Y}\right)=0.9922779<C\left(\vec{X}',\vec{Y}'\right)=0.9988681$$

$$E\left(\overrightarrow{X},\overrightarrow{Y}\right)=0.969697<E\left(\overrightarrow{X}',\overrightarrow{Y}'\right)=0.9767442$$

$$J\left(\overrightarrow{X},\overrightarrow{Y}\right)=0.9411765<J\left(\overrightarrow{X}',\overrightarrow{Y}'\right)=0.9545455$$

$$O_1\left(\overrightarrow{X},\overrightarrow{Y}\right)=1.2307692<O_1\left(\overrightarrow{X}',\overrightarrow{Y}'\right)=01.2352941$$

$$O_2\left(\overrightarrow{X},\overrightarrow{Y}\right)=0.8<O_2\left(\overrightarrow{X}',\overrightarrow{Y}'\right)=0.8076923$$

We claim that the fact that these measures do not follow the transfer principle is a good property: in no way are $\overrightarrow{X}'$ and $\overrightarrow{Y}'$ less similar than $\overrightarrow{X}$ and $\overrightarrow{Y}$. Indeed, the last coordinates in $\overrightarrow{X}$ and $\overrightarrow{Y}$ and in $\overrightarrow{X}'$ and $\overrightarrow{Y}'$ are equal and in the first coordinates, the difference between the coordinates is one but the first coordinates in $\overrightarrow{X}'$ and $\overrightarrow{Y}'$ are larger than in $\overrightarrow{X}$ and $\overrightarrow{Y}$ so that, intuitively, $\overrightarrow{X}'$ and $\overrightarrow{Y}'$ are "more similar" than $\overrightarrow{X}$ and $\overrightarrow{Y}$ which is also expressed by the values of the similarity measures C, E, J, O_1 and O_2.

This also shows that similarity measures, in addition to the fact that they act on two vectors (while concentration measures act on one vector), do not follow (and do not have to follow) the reverse concentration properties (also called dispersion properties).

III. Another good property for similarity measures

Let again $\overrightarrow{X}=\left(x_1,...,x_N\right)$ and $\overrightarrow{Y}=\left(y_1,...,y_N\right)$ be two vectors with $x_i,y_i\geq 0$ for all $i=1,...,N$ and such that $\overrightarrow{X}\neq\overrightarrow{Y}$. It is then logic that, if we add one of these vectors (supposed to be non-zero) to these two vectors, then the similarity should increase strictly. So if we denote by sim any similarity measure, we require

$$sim\left(\overrightarrow{X}+\overrightarrow{Y},2\overrightarrow{Y}\right)>sim\left(\overrightarrow{X},\overrightarrow{Y}\right) \tag{23}$$

and

$$sim\left(2\vec{X}, \vec{X}+\vec{Y}\right) > sim\left(\vec{X}, \vec{Y}\right) \tag{24}$$

The Cosine satisfies this property as the next Theorem shows.

Theorem III.1: For all $\vec{X}$, $\vec{Y}$ as above, such that one vector is not a multiple of the other, we have

$$C\left(\vec{X}+\vec{Y}, 2\vec{Y}\right) > C\left(\vec{X}, \vec{Y}\right) \tag{25}$$

and

$$C\left(2\vec{X}, \vec{X}+\vec{Y}\right) > C\left(\vec{X}, \vec{Y}\right) \tag{26}$$

Proof: It suffices to prove (25); (26) follows from (25) upon interchange of $\vec{X}$ and $\vec{Y}$ and the fact that the cosine is symmetric. We have to show that

$$\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} < \frac{\sum (x_i + y_i) 2y_i}{\sqrt{\sum (x_i + y_i)^2}\sqrt{\sum (2y_i)^2}} \tag{27}$$

This is equivalent with

$$\left(\sum x_i y_i\right)\sqrt{\sum (x_i + y_i)^2} < \left(\sum (x_i + y_i) y_i\right)\sqrt{\sum x_i^2}$$

or

$$\left(\sum x_i y_i\right)^2 \left(\sum x_i^2 + 2\sum x_i y_i + \sum y_i^2\right) < \left(\left(\sum x_i y_i\right)^2 + 2\left(\sum x_i y_i\right)\left(\sum y_i^2\right) + \left(\sum y_i^2\right)^2\right)\left(\sum x_i^2\right)$$

or

$$2\left(\sum x_i y_i\right)^3 + \left(\sum x_i y_i\right)^2 \left(\sum y_i^2\right) < 2\left(\sum x_i y_i\right)\left(\sum x_i^2\right)\left(\sum y_i^2\right) + \left(\sum x_i^2\right)\left(\sum y_i^2\right)^2 \qquad (28)$$

Since $\vec{X}$ and $\vec{Y}$ are not multiples we have

$$\left(\sum x_i y_i\right)^2 < \left(\sum x_i^2\right)\left(\sum y_i^2\right) \qquad (29)$$

, hence

$$2\left(\sum x_i y_i\right)^3 + \left(\sum x_i y_i\right)^2 \left(\sum y_i^2\right) < 2\left(\sum x_i y_i\right)\left(\sum x_i^2\right)\left(\sum y_i^2\right) + \left(\sum x_i^2\right)\left(\sum y_i^2\right)^2$$

which proves (28) and hence (25). If $\vec{X}$ and $\vec{Y}$ are multiples we have that (25) and (26) are equalities. □

The above proof essentially showed that the angle between $\vec{X} + \vec{Y}$ and $2\vec{Y}$ (or $\vec{Y}$) is smaller than the one between $\vec{X}$ and $\vec{Y}$.

Also Dice satisfies this property. In fact, L. Waltman proved that Dice even satisfies the following stronger property : let $\vec{Z} = (z_1, ..., z_N) \neq \vec{O}$ be any vector with positive (or zero) coordinates. Then (denoting by sim a similarity measure)

$$\text{sim}\,(\vec{X} + \vec{Z}\,,\ \vec{Y} + \vec{Z}\,) > \text{sim}\,(\vec{X}\,, \vec{Y}\,) \qquad (30)$$

We have the following result of L. Waltman.

Theorem III.2 (L. Waltman) : For any vectors $\vec{X}, \vec{Y}$, $\vec{X} \neq \vec{Y}$, and any nonzero vector $\vec{Z}$ as above we have

$$\text{E}\,(\vec{X} + \vec{Z}\,,\ \vec{Y} + \vec{Z}\,) > \text{E}(\vec{X}\,, \vec{Y}\,) \qquad (31)$$

Proof :

$$\mathrm{E}(\vec{X}+\vec{Z}\,,\,\vec{Y}+\vec{Z}\,)$$

$$= \frac{2\sum(x_i+z_i)(y_i+z_i)}{\sum(x_i+z_i)^2+\sum(y_i+z_i)^2}$$

$$= \frac{2\sum x_i y_i + 2\sum z_i(x_i+y_i+z_i)}{\sum x_i^2+\sum z_i(2x_i+z_i)+\sum y_i^2+\sum z_i(2y_i+z_i)}$$

$$= \frac{2\sum x_i, y_i + 2\sum z_i\,(x_i+y_i+z_i)}{\sum x_i^2+\sum y_i^2+2\sum z_i\,(x_i+y_i+z_i)}$$

$$> \frac{2\sum x_i y_i}{\sum x_i^2+\sum y_i^2} = \mathrm{E}\left(\vec{X},\vec{Y}\right)$$

since $\vec{Z}\neq\vec{O}$ and since $\vec{X}\neq\vec{Y}$. □

Note that the property studied in this section follows by taking $\vec{Z}=\vec{X}$ or $\vec{Z}=\vec{Y}$. Also the principle of nominal increase is included in the above general property, by taking $\vec{Z}=\vec{A}$.

Since Jaccard's index has a strickly increasing relation with Dice (see (3) or (11), it follows that also J satisfies this general property. By the results on the principle of nominal increase, it follows that none of the other measures studied here satisfy this general principle.

However we have the following Theorem on O_2 showing that the properties (23) and (24) are always is true.

Theorem III.3: Let $\vec{X}$ and $\vec{Y}$ be any vectors such that $\vec{X}\neq\vec{Y}$. Then

$$O_2\left(\vec{X}+\vec{Y},2\vec{Y}\right) > O_2\left(\vec{X},\vec{Y}\right) \tag{32}$$

and

$$O_2\left(2\vec{X}, \vec{X}+\vec{Y}\right) > O_2\left(\vec{X}, \vec{Y}\right) \tag{33}$$

Proof: The proof needs several parts. Since we want to prove both (32) and (33) we cannot assume that $\sum x_i^2 \le \sum y_i^2$ (or $\ge$). So we have to deal with all cases.

(i) Let

$$\sum x_i^2 \le \sum y_i^2 \tag{34}$$

and let $\vec{X}$ and $\vec{Y}$ not be multiples. Then

$$\sum \left(x_i + y_i\right)^2$$

$$= \sum x_i^2 + 2\sum x_i y_i + \sum y_i^2$$

$$\le \sum y_i^2 + 2\sqrt{\sum x_i^2}\sqrt{\sum y_i^2} + \sum y_i^2$$

$$\le 4\sum y_i^2$$

by the Cauchy-Schwarz inequality and by (34).

So

$$O_2\left(\vec{X}+\vec{Y}, 2\vec{Y}\right) = \frac{\sum \left(x_i + y_i\right) 2y_i}{4\sum y_i^2}$$

and hence we have to show, for (32)

$$\frac{\sum x_i y_i}{\sum y_i^2} < \frac{\sum \left(x_i + y_i\right) 2y_i}{4\sum y_i^2} \tag{35}$$

again using (34). This is equivalent with

$$4\sum x_i y_i < 2\sum x_i y_i + 2\sum y_i^2$$

which is true by (29) since $\vec{X}$ and $\vec{Y}$ are not multiples and by (34). This proves (32) in case we have (34).

(ii) Let now

$$\sum x_i^2 > \sum y_i^2 \tag{36}$$

and $\vec{X}$, $\vec{Y}$ are not multiples.

By definition of O_2 we have to show that

$$\frac{\sum x_i y_i}{\sum x_i^2} < \frac{\sum (x_i + y_i) 2y_i}{\max\left(\sum (x_i + y_i)^2, 4\sum y_i^2\right)} \tag{37}$$

Now we cannot know which of the numbers $\sum (x_i + y_i)^2$ and $4\sum y_i^2$ is the largest. So we have to prove both inequalities (38) and (39):

$$\frac{\sum x_i y_i}{\sum x_i^2} < \frac{\sum (x_i + y_i) 2y_i}{\sum (x_i + y_i)^2} \tag{38}$$

$$\frac{\sum x_i y_i}{\sum x_i^2} < \frac{\sum (x_i + y_i) 2y_i}{4\sum y_i^2} \tag{39}$$

Now (38) is valid if and only if

$$\left(\sum x_i y_i\right)\left(\sum (x_i + y_i)^2\right) < \left(\sum x_i^2\right)\left(2\sum (x_i + y_i) y_i\right)$$

or

$$2\left(\sum x_i y_i\right)^2 + \left(\sum x_i y_i\right)\left(\sum y_i^2\right) < \left(\sum x_i y_i\right)\left(\sum x_i^2\right) + 2\left(\sum x_i^2\right)\left(\sum y_i^2\right)$$

but this is valid by (29) (since $\vec{X}$ and $\vec{Y}$ are not multiples) and by (36).

Now (39) is valid if and only if

$$4\left(\sum x_i y_i\right)\left(\sum y_i^2\right) < \left(\sum x_i^2\right)\left(2\sum x_i y_i + 2\sum y_i^2\right)$$

But

$$2\left(\sum x_i y_i\right)\left(\sum y_i^2\right) < 2\left(\sum x_i y_i\right)\left(\sum x_i^2\right)$$

by (36) and

$$2\left(\sum x_i y_i\right)\left(\sum y_i^2\right)$$

$$< 2\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}\left(\sum y_i^2\right)$$

$$< 2\left(\sum x_i^2\right)\left(\sum y_i^2\right)$$

by (29) and (36).

Hence we have proved (34) completely and (35) follows in the same way if $\vec{X}$ and $\vec{Y}$ are not multiples.

(iii) Let now $\vec{X} = \alpha\vec{Y}$ for a certain $\alpha > 0$, $\alpha \neq 1$. We have

$$O_2\left(\vec{X},\vec{Y}\right)=\frac{\alpha}{\max\left(\alpha^2,1\right)} \tag{40}$$

$$O_2\left(\vec{X}+\vec{Y},2\vec{Y}\right)=\frac{2\left(\alpha+1\right)}{\max\left(\left(\alpha+1\right)^2,4\right)} \tag{41}$$

Hence for (32) we have to prove

$$\frac{2\left(\alpha+1\right)}{\max\left(\left(\alpha+1\right)^2,4\right)}>\frac{\alpha}{\max\left(\alpha^2,1\right)} \tag{42}$$

(I) Let $\left(\alpha+1\right)^2\geq 4$ (this is so if and only if $\alpha\geq 1$). Then (42) boils down to

$$\frac{2\left(\alpha+1\right)}{\left(\alpha+1\right)^2}>\frac{1}{\alpha}$$

which is so since $\alpha\geq 1$ and $\alpha\neq 1$ (since $\vec{X}\neq\vec{Y}$).

(II) Let $\left(\alpha+1\right)^2<4$ (this is so if and only if $\alpha<1$). Then (42) boils down to

$$\frac{\alpha+1}{2}>\alpha$$

which is so since $\alpha<1$.

This completes the proof of (32) and the one of (33) is similar. □

For the overlap measure O_1, the results are completely negative with respect to the property studied in this section. A simple example proves this: take $\vec{X}=\left(2,1\right)$, $\vec{Y}=\left(3,1\right)$. Then

$O_1 = \frac{7}{5}$. But $\vec{X} + \vec{Y} = (5,2)$ and $2\vec{Y} = (6,2)$ yielding $O_1\left(\vec{X} + \vec{Y}, 2\vec{Y}\right) = \frac{34}{29} < O_1$. Also:

$O_1\left(2\vec{X}, \vec{X} + \vec{Y}\right) = \frac{24}{20} < O_1$.

We can conclude that, with respect to the good similarity properties studied in Section II and this section, the overlap measure O_2 outperforms the overlap measure O_1, although, for the principle of nominal increase, we only have partial positive results for O_2 (but O_1 scores negative in both cases).

We close this section by studying a variant of the property (23) (or (24)) studied in this section. Let us focus on (23). Instead of adding the vector $\vec{Y}$ to both vectors $\vec{X}$ and $\vec{Y}$, we simply add the vector $\vec{Y}$ to $\vec{X}$ and keep $\vec{Y}$ as second vector. Then we require a similarity measure sim to satisfy (variant of inequality (23)):

$$sim\left(\vec{X} + \vec{Y}, \vec{Y}\right) > sim\left(\vec{X}, \vec{Y}\right) \tag{43}$$

(and similarly for the variant of inequality (24)). What similarity measures satisfy this property ?

Let us start with a simple remark: Cosine satisfies this property (if $\vec{X}$ and $\vec{Y}$ are not multiples; otherwise we have equal values). This, trivially, follows from Theorem III.1 and the fact that

$$C\left(\vec{X} + \vec{Y}, \vec{Y}\right) = C\left(\vec{X} + \vec{Y}, 2\vec{Y}\right)$$

, using formula (4).

For Dice's measure E we have a double counterexample: Let $\vec{X} = (2,1)$, $\vec{Y} = (3,1)$, then

$$E\left(\vec{X} + \vec{Y}, \vec{Y}\right) = \frac{34}{39} < E\left(\vec{X}, \vec{Y}\right) = \frac{14}{15}$$

and

$$E\left(\vec{X},\vec{X}+\vec{Y}\right)=\frac{24}{34}<E\left(\vec{X},\vec{Y}\right).$$

By formula (11) (the strictly increasing relationship between J (Jaccard's index) and E) we also have that the above counterexamples for E for this variant property are also true for J.

For overlap measure O_2 we also have a double counterexample for the same vectors $\vec{X}=(2,1)$ and $\vec{Y}=(3,1)$ above:

$$O_2\left(\vec{X}+\vec{Y},\vec{Y}\right)=\frac{17}{29}<O_2\left(\vec{X},\vec{Y}\right)=\frac{7}{10}$$

and

$$O_2\left(\vec{X},\vec{X}+\vec{Y}\right)=\frac{12}{29}<O_2\left(\vec{X},\vec{Y}\right).$$

For overlap measure O_1, there is a "50%" result, stated and proved in the next Proposition.

Proposition III.4: For all vectors $\vec{X}$ and $\vec{Y}$ we have

(i) $\quad \sum y_i^2 \le \sum x_i^2 \Rightarrow O_1\left(\vec{X}+\vec{Y},\vec{Y}\right)>O_1\left(\vec{X},\vec{Y}\right)$

(ii) $\quad \sum y_i^2 \ge \sum x_i^2 \Rightarrow O_1\left(\vec{X},\vec{X}+\vec{Y}\right)>O_1\left(\vec{X},\vec{Y}\right)$

and there are counterexamples in the other cases.

Proof: It suffices to prove (i).

$$O_1\left(\vec{X},\vec{Y}\right) < O_1\left(\vec{X}+\vec{Y},\vec{Y}\right)$$

if and only if

$$\frac{\sum x_i y_i}{\min\left(\sum x_i^2, \sum y_i^2\right)} < \frac{\sum\left(x_i + y_i\right) y_i}{\min\left(\sum\left(x_i + y_i\right)^2, \sum y_i^2\right)} = \frac{\sum\left(x_i + y_i\right) y_i}{\sum y_i^2}$$

or

$$\left(\sum x_i y_i\right)\left(\sum y_i^2\right) < \min\left(\sum x_i^2, \sum y_i^2\right)\left(\sum\left(x_i + y_i\right) y_i\right)$$

or

$$\left(\sum x_i y_i\right)\left(\sum y_i^2\right) < \min\left(\sum x_i^2, \sum y_i^2\right)\left(\sum x_i y_i + \sum y_i^2\right) \tag{44}$$

So if $\sum y_i^2 \le \sum x_i^2$, the (44) is trivially satisfied. Similarly we can prove (ii).

To provide a counterexample for (i) in case $\sum y_i^2 > \sum x_i^2$, we first try to prove (i) in this case: by (44) we have to prove:

$$\left(\sum x_i y_i\right)\left(\sum y_i^2\right) < \left(\sum x_i^2\right)\left(\sum x_i y_i + \sum y_i^2\right)$$

or

$$\left(\sum x_i y_i\right)\left(\sum y_i^2 - \sum x_i^2\right) < \sum x_i^2 \sum y_i^2$$

This is not possible: we will make an example for which we have

$$\left(\sum x_i y_i\right)\left(\sum y_i^2 - \sum x_i^2\right) > \left(\sum x_i^2\right)\left(\sum y_i^2\right)$$

as follows: Let $\vec{X} = (2,1)$, $\vec{Y} = (1,5)$. Then $\sum y_i^2 > \sum x_i^2$ and

$$\left(\sum x_i y_i\right)\left(\sum y_i^2 - \sum x_i^2\right) = 7(26-5) = 147 > \left(\sum x_i^2\right)\left(\sum y_i^2\right) = 130.$$

We now know that we have a counterexample for (i) in Proposition III.4:

$$O_1\left(\vec{X}+\vec{Y},\vec{Y}\right) = \frac{33}{26} < O_1\left(\vec{X},\vec{Y}\right) = \frac{7}{5}$$

Similarly for (ii). □

Corollary III.5: For all vectors $\vec{X}$ and $\vec{Y}$ we have, if $\sum x_i^2 = \sum y_i^2$ that

$$O_1\left(\vec{X}+\vec{Y},\vec{Y}\right) > O_1\left(\vec{X},\vec{Y}\right)$$

and

$$O_1\left(\vec{X},\vec{X}+\vec{Y}\right) > O_1\left(\vec{X},\vec{Y}\right)$$

Proof: This follows trivially from Proposition III.4. □

We conclude that, except for C and "50%" for O_1, none of the similarity measures studied here satisfy the variant property as studied in this section. This is due to the fact that, when taking $\vec{X}+\vec{Y}$, we still used $\vec{Y}$ and not $2\vec{Y}$ as in the original property in this section. Although all used measures are normalized, this, apparently, makes a difference.

IV. Conclusions, remarks and suggestions for further research

The principle of nominal increase, defined in econometrics, was tested on similarity measures, hence in a two-dimensional framework. We proved that Dice's measure and Jaccard's index satisfy this logical property for similarity measures.

Cosine does not satisfy this principle (by example). Since $O_1O_2 = C^2$, at least one of the overlap measures O_1 and O_2 do not satisfy this principle either. It turns out that both O_1 and O_2 do not satisfy it while O_2 satisfies this principle in case the vectors $\vec{X}$ and $\vec{Y}$ are multiples or in case the L^1-norms of $\vec{X}$ and $\vec{Y}$ are equal.
We remark that the "stronger" property of concentration theory, the transfer principle, cannot be used in the context of similarity measures.

Next we claim that, adding one of the two vectors to the two vectors should increase the similarity. Now we show that all measures (except O_1) satisfy this principle. This and the previous principle show that O_2 outperforms O_1 with respect to these logical principles.

A variant of the latter principle (only adding one of the vectors to the other one) is also studied. Cosine satisfies this principle while Dice's measure, Jaccard's index and O_2 do not. For O_1 we have positive results in case one L^2-norm is larger than or equal to the other L^2-norm, but also counterexamples are given for O_1.

In van Eck and Waltman (2009) one presents the Association Strength, denoted by S:

$$S = \frac{\sum x_i y_i}{\sum x_i^2 \sum y_i^2} \tag{45}$$

This measure, however, does not satisfy any of the properties studied above. Indeed, for the principle of nominal increase, take $\vec{X} = (3,1)$, $\vec{Y} = (6,2)$, $\vec{A} = (1,1)$. Then $S(\vec{X},\vec{Y}) = 0.05 > S(\vec{X}+A,\vec{Y}+A) = 0.0293103$.

For the second principle (adding one of the two vectors to both vectors), we take $\vec{X} = (2,1)$, $\vec{Y} = (3,1)$, $\vec{X}+\vec{Y} = (5,2)$, $2\vec{Y} = (6,2)$ and $S(\vec{X}+\vec{Y},2\vec{Y}) = 0.0293103 < S(\vec{X},\vec{Y}) = 0.14$. For the variant we have $S(\vec{X}+\vec{Y},\vec{Y}) = 0.0586207 < S(\vec{X},\vec{Y}) = 0.14$. Note also the very small values of S in all these examples. Although actual values of a similarity measure are not so important, we do not think it is logical to have a similarity value of 0.14 for the vectors $\vec{X} = (2,1)$ and $\vec{Y} = (3,1)$ (and similar for the other examples). These low values are due to the division in (45) of $\sum x_i y_i$ by $\sum x_i^2 \sum y_i^2$ and not by their square roots as is the case for Cosine.

All these results can be summarized in Table 1 (Y = yes, N = no, P = partially true

Table 1. Scores of similarity measures with respect to three logical properties

Results	E	J	C	O_1	O_2	S
$(\vec{X}+\vec{A},\vec{Y}+\vec{A})$	Y	Y	N	N	NP	N
$(\vec{X}+\vec{Y},2\vec{Y})$	Y	Y	Y	N	Y	N
$(\vec{X}+\vec{Y},\vec{Y})$	N	N	Y	P	N	N

This Table clearly shows the complementarity of similarity measures with respect to the studied properties: we do not have (at least in our paper) a similarity measure that satisfies all properties.

In van Eck and Waltman (2009) one finds a common generalization of several of the similarity measures studied here: the measure G_p, dependent on a parameter $p \in \mathbb{R} \setminus \{0\}$ (the real numbers without 0).

$$G_p = \frac{2^{\frac{1}{p}} \sum x_i y_i}{\left(\left(\sum x_i^2 \right)^p + \left(\sum y_i^2 \right)^p \right)^{\frac{1}{p}}} \tag{46}$$

They indicate the following results

$$\lim_{p \to -\infty} G_p = O_1 \tag{47}$$

$$\lim_{p \to 0} G_p = C \tag{48}$$

$$G_1 = E \tag{49}$$

$$\lim_{p \to +\infty} G_p = O_2 \tag{50}$$

So, from our above results, whether or not G_p satisfies one or more of the good properties for similarity measures, depends on the parameter p. It is a difficult open problem to characterise p such that none, one, two, three or all good properties studied here are valid for G_p (or its limits). Of course, also other similarity measures can be studied in this context.

References

P. Ahlgren, B. Jarneving and R. Rousseau (2003). 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.

B.R. Boyce, C.T. Meadow and D.H. Kraft (1995). Measurement in Information Science. Academic Press, New York, USA.

L. Egghe (2005). Power Laws in the Information Production Process: Lokaian Informetrics. Elsevier, Oxford, UK.

L. Egghe (2009). New relations between similarity measures for vectors based on vector norms. Journal of the American Society for Information Science and Technology 60(2), 232-239.

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

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

L. Egghe and R. Rousseau (1990). Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier, Amsterdam, the Netherlands.

D.A. Grossman and O. Freider (1998). Information Retrieval Algorithms and heuristics. Kluwer, Boston, USA.

R.M. Losee (1998). Text Retrieval and Filtering: Analytical Models of Performance. Kluwer, Boston, USA.

G. Salton and M.J. McGill (1987). Introduction to modern Information Retrieval. McGraw-Hill, New York, USA.

J. Tague-Sutcliffe (1995). Measuring Information: An Information Services Perspective. Academic Press, New York, USA.

C.J. van Rijsbergen (1979). Information Retrieval. Butterworths, London, UK.

N.J. van Eck and L. Waltman (2009). How to normalize cooccurrence data ? An analysis of some well-known similarity measures. Journal of the American Society for Information Science and Technology 60(8), 1635-1651.