New relations between similarity measures for vectors based on vector norms

New relations between similarity measures for vectors based on vector norms

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]

and

Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium

leo.egghe@uhasselt.be

ABSTRACT

The well-known similarity measures Jaccard, Salton's cosine, Dice and several related overlap measures for vectors are compared. While general relations are not possible to prove, we study these measures on the "trajectories" of the form $\left\|\vec{X}\right\| = a\left\|\vec{Y}\right\|$, where $a > 0$ is a constant and $\|\cdot\|$ denotes the Euclidean norm of a vector. In this case, direct functional relations between these measures are proved. For Jaccard we prove that it is a convexly increasing function of Salton's cosine measure, but always smaller than or equal to the latter, hereby explaining a curve, experimentally found by Leydesdorff. All the other measures have a linear relation with Salton's cosine, reducing even to equality, in case $a = 1$. Hence for

---

[1] Permanent address

Key words and phrases: similarity measure, Jaccard, Salton's cosine measure, Dice, overlap measure

equally normed vectors (e.g. for normalized vectors) we, essentially, only have Jaccard's measure and Salton's cosine measure, since all the other measures are equal to the latter.

I. Introduction

The similarity measures Jaccard, Salton's cosine (briefly cosine), Dice and related overlap measures are best known in their set-theoretic version. Let us repeat the well known definitions which can be found e.g. in the classical monographs Boyce, Meadow and Kraft (1995), Tague-Sutcliffe (1995), Grossman and Frieder (1998), Losee (1988), Salton and McGill (1987) and Van Rijsbergen (1979) and see also Egghe and Michel (2002, 2003).

In general one has a universe Ω from which subsets A, B,… are considered as e.g. in information retrieval (IR) where sets A, B,… are document sets retrieved from queries that were put in an IR system for which Ω is the entire database. A good similarity measure S measures the degree of similarity between any two subsets A, B of Ω and hence is a function

$$S: \Omega \times \Omega \to \mathbb{R}^+$$
$$(A, B) \to S(A, B) \tag{1}$$

where some elementary requirements must be fulfilled: $S(A, B)$ should be minimal (say 0 in most cases) if $A \cap B = \varnothing$ and should be maximal (say 1 in most cases) if $A = B \neq \varnothing$ (in which case S ranges in the interval $[0,1]$). Let us for any subset A, B of Ω denote by $|A|, |B|$ the cardinality of A respectively B, i.e. the number of elements in A, B.

Jaccard's measure, denoted J, is a symmetric overlap measure defined as follows, for A, B subsets of Ω, $A, B \neq \varnothing$.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Salton's cosine measure, denoted Cos is also symmetrical and is defined as

$$Cos(A, B) = \frac{|A \cap B|}{\sqrt{|A||B|}} \tag{3}$$

(note that $\sqrt{|A||B|}$ is the geometric average of $|A|$ and $|B|$). Why this measure is called Cosine is well-known but will be repeated further on.

Dice's measure, denoted E, is also a symmetric similarity measure and is defined as

$$E(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{4}$$

(note that now we use the arithmetic average $\frac{1}{2}(|A| + |B|)$ of $|A|$ and $|B|$). Sometimes one considers the generalized Dice measure E_α, for $\alpha \in]0,1[$

$$E_\alpha(A,B) = \frac{|A \cap B|}{\alpha|A| + (1-\alpha)|B|} \tag{5}$$

Note that $E = E_{\frac{1}{2}}$; the general formula (5) uses a general convex combination of $|A|$ and $|B|$.

A symmetrical measure, for which we do not have a name, but denoted by N is the following

$$N(A,B) = \sqrt{2}\frac{|A \cap B|}{\sqrt{|A|^2 + |B|^2}} \tag{6}$$

Two other symmetrical overlap measures are

$$O_1(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} \tag{7}$$

$$O_2(A,B) = \frac{|A \cap B|}{\max(|A|,|B|)} \tag{8}$$

Elementary calculations yield the following general inequalities between these measures

$$J \le O_2 \le E \le N \le Cos \le O_1 \tag{9}$$

Two non-symmetrical overlap measures are

$$P = \frac{|A \cap B|}{|A|} \tag{10}$$

and

$$R = \frac{|A \cap B|}{|B|} \tag{11}$$

In IR, P is the precision if A = ret (the set of retrieved documents) and B = rel (the set of relevant documents) in which case R is called recall. Other interpretations of (10) and (11) in terms of fallout and miss can also be given – see Egghe (2007, 2008). Note that E respectively E_α are the harmonic, respectively generalized harmonic mean of P and R:

$$E_\alpha = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}} \tag{12}$$

and $E = E_{\frac{1}{2}}$.

Further inequalities are (elementary calculation)

$$O_2 \le P, R, E_\alpha \le O_1 \tag{13}$$

In Egghe and Michel (2002) it is explained how to interpret the above definitions in terms of similarity measures between two vectors $\vec{X} = (x_1, x_2, ..., x_n), \vec{Y} = (y_1, y_2, ..., y_n)$ where $n \in \mathbb{N}$. In IR, $\vec{X}$ and $\vec{Y}$ can describe queries and documents where e.g. $\vec{X}$ is a query and $\vec{Y}$ is a document and where x_i or $y_i = 1$ if key word i appears in the query respectively the document and where x_i or $y_i = 0$ if this is not the case. Of course $\vec{X}$ and $\vec{Y}$ can both represent query vectors or document vectors in which case (e.g. the latter one) similarity between two documents is measured.

The step from the above set-theoretic similarity measures to similarity measures for vectors is taken as follows. Denote

$$A = \{i \in \{1, ..., n\} \mid\mid x_i = 1\} \tag{14}$$

$$B = \{i \in \{1, ..., n\} \mid\mid y_i = 1\} \tag{15}$$

then

$$|A \cap B| = \vec{X} \cdot \vec{Y} = \sum_{i=1}^{n} x_i \cdot y_i \tag{16}$$

where $\vec{X} \cdot \vec{Y}$ is the inproduct (sometimes also denoted by $< \vec{X}, \vec{Y} >$; the inproduct is also called the dot product or inner product) of the vectors $\vec{X}$ and $\vec{Y}$. Also

$$|A| = \sum_{i=1}^{n} x_i^2 = \left\|\vec{X}\right\|_2^2 \tag{17}$$

$$|B| = \sum_{i=1}^{n} y_i^2 = \left\|\vec{Y}\right\|_2^2 \tag{18}$$

, the squares of the Euclidean norms of the vectors $\vec{X}$ and $\vec{Y}$. Finally, we also have

$$|A \cup B| = \left\|\vec{X}\right\|_2^2 + \left\|\vec{Y}\right\|_2^2 - \vec{X} \cdot \vec{Y} \tag{19}$$

In this way, all the above similarity measures (for sets) can be redefined for vectors $\vec{X} = (x_1, x_2, ..., x_n), \vec{Y} = (y_1, y_2, ..., y_n)$ and where we immediately extend the defenition to the non-binary case: all coordinates are positive real numbers: $x_i, y_i \in \mathbb{R}^+$ for all $i = 1, ..., n$. Since we extensively need these formulae in the sequel we will define them here explicitely.

$$J = J(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{\left\|\vec{X}\right\|^2 + \left\|\vec{Y}\right\|^2 - \vec{X} \cdot \vec{Y}}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} x_i y_i} \tag{20}$$

where we denote, for the sake of simplicity, $\|\cdot\| = \|\cdot\|_2$.

$$Cos = Cos(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{\left\|\vec{X}\right\| \cdot \left\|\vec{Y}\right\|}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{21}$$

Now it is clear why Cos is called the cosine measure: $Cos(\vec{X}, \vec{Y})$ is indeed the cosine of the angle between the vectors $\vec{X}$ and $\vec{Y}$. Note also that, if $\left\|\vec{X}\right\| = \left\|\vec{Y}\right\| = 1$, the cosine $Cos(\vec{X}, \vec{Y})$ equals the simple dot product $\vec{X}.\vec{Y}$. Further

$$E = E(\vec{X}, \vec{Y}) = \frac{2\vec{X} \cdot \vec{Y}}{\left\|\vec{X}\right\|^2 + \left\|\vec{Y}\right\|^2}$$

$$= \frac{2\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2} \tag{22}$$

and in the generalized form

$$E_\alpha = E_\alpha(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{\alpha \left\|\vec{X}\right\|^2 + (1-\alpha)\left\|\vec{Y}\right\|^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\alpha \sum_{i=1}^{n} x_i^2 + (1-\alpha)\sum_{i=1}^{n} y_i^2} \tag{23}$$

Further

$$N = N(\vec{X}, \vec{Y}) = \sqrt{2}\, \frac{\vec{X} \cdot \vec{Y}}{\sqrt{\left\|\vec{X}\right\|^4 + \left\|\vec{Y}\right\|^4}}$$

$$= \sqrt{2}\, \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)^2 + \left(\sum_{i=1}^{n} y_i^2\right)^2}} \tag{24}$$

$$O_1 = O_1(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{\min\left(\left\|\vec{X}\right\|^2, \left\|\vec{Y}\right\|^2\right)}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\min\left(\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} y_i^2\right)} \tag{25}$$

$$O_2 = O_2(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{\max\left(\left\|\vec{X}\right\|^2, \left\|\vec{Y}\right\|^2\right)}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\max\left(\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} y_i^2\right)} \tag{26}$$

and, finally,

$$P = P(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{\left\|\vec{X}\right\|^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{27}$$

$$R = R(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{\left\|\vec{Y}\right\|^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^2} \tag{28}$$

It is clear that general functional relations, using general sums $\sum_{i=1}^{n}$ between these vectors are not possible. Yet if one looks at Fig.3 in Leydesdorff (2008), it is clear that there is an almost pure functional relationship between J and C – see Fig.1 below which is a reproduction of this graph (with kind permission from Wiley Interscience).

The cloud of points hardly has any thickness, expressing that there should be a functional relationship between J and C. The study of this relation is executed in the next section. We will show that, in all cases where $\left\|\vec{X}\right\| = a\left\|\vec{Y}\right\|$ ($a > 0$, a constant) we have a concavely increasing function as in Fig.1. Also, for $a = 1$ we show that the relation is (denote Cos = C)

$$J = \frac{C}{2 - C} \tag{29}$$

and we show that this function almost exactly fits the points in Fig.1. We also show that all functional relations between J and C are below the first bissectrix and that (29) is the highest curve for all $a > 0$. We also give a necessary and sufficient condition for one curve to be above another one, in terms of the a-values. We also present a formula for calculating the distance between two such curves and estimate from Fig.1 that, in this dataset, if $\left\|\vec{Y}\right\| > \left\|\vec{X}\right\|$, then $\left\|\vec{Y}\right\| \le 2\left\|\vec{X}\right\|$ (hence $\left\|\vec{X}\right\| \ge 0.5\left\|\vec{Y}\right\|$) limiting the possible a-values to $a \in [0.5, 2]$. This fully explains the “sharp” graph in Fig.1

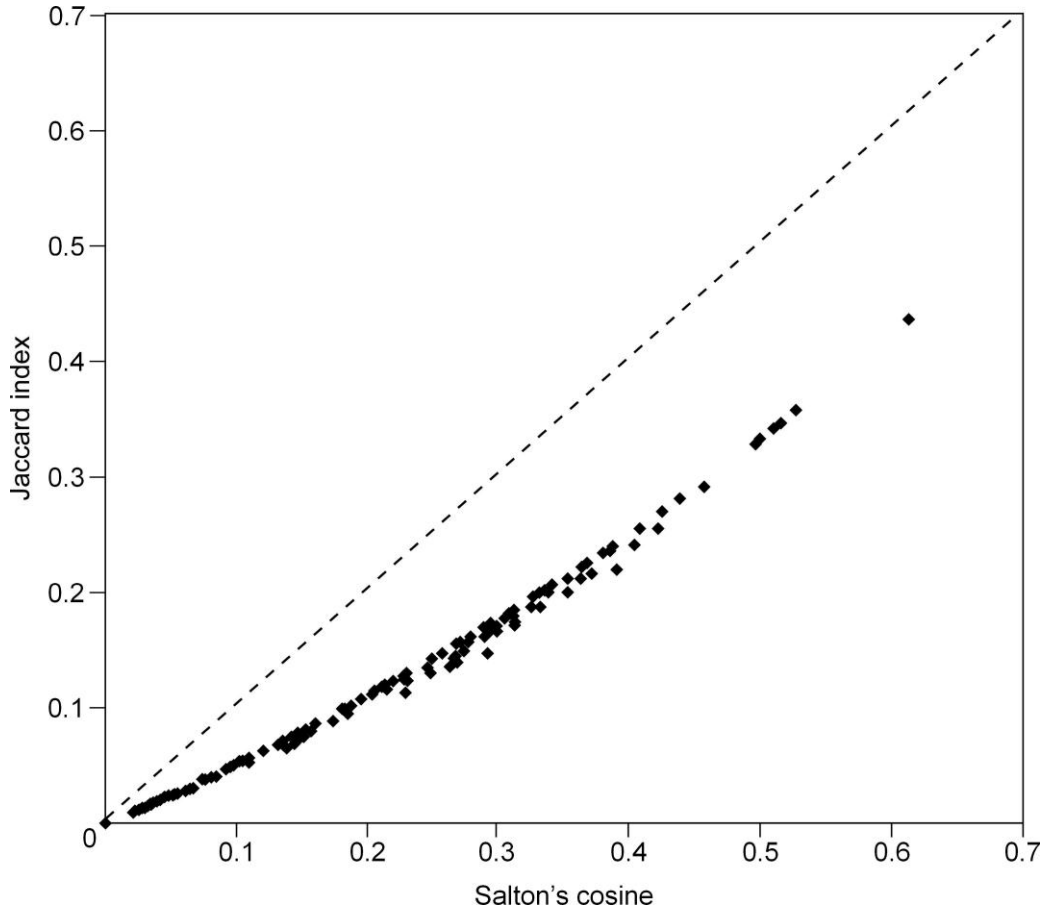

Fig.1 Relation between J and Cos: case of 24 information scientists, hence (24 x 23)/2 points dealing with citation data of Ahlgren, Jarneving and Rousseau (2003). Reprinted with kind permission from Willey.

The third section is devoted to the relation of the other similarity measures with $\mathrm{Cos} = \mathrm{C}$. There we show that all relations are linear in C and even that all measures are equal with C for $a = 1$ (i.e. on $\left\|\vec{X}\right\| = \left\|\vec{Y}\right\|$). All straight lines (except two) are below the first bissectrix (being the line of equality between one measure and C, which is the case for $a = 1$, as mentioned above).

The paper ends with some suggestions for further research and some open problems.

II. The relation between Jaccard and Salton's Cosine measure

Let $\vec{X} = (x_1, ..., x_n), \vec{Y} = (y_1, ..., y_n)$ be non-zero vectors such that $\vec{X} \cdot \vec{Y} \neq 0$. This means that $\vec{X}$ and $\vec{Y}$ are not perpendicular (otherwise all similarity measures are zero in which case no further comparison is necessary). We first prove a Lemma.

Lemma II.1:

Denoting $Cos(\vec{X}, \vec{Y})$ by C we generally have for all vectors $\vec{X}$ and $\vec{Y}$

$$J = J(\vec{X}, \vec{Y}) = \frac{C}{\sqrt{\frac{\sum x_i^2}{\sum y_i^2}} + \sqrt{\frac{\sum y_i^2}{\sum x_i^2}} - C} \tag{30}$$

and where we denote, henceforth $\sum_{i=1}^{n} x_i^2 = \sum x_i^2$ and $\sum_{i=1}^{n} y_i^2 = \sum x_i^2$.

Proof:

Denoting also

$$\sum_{i=1}^{n} x_i y_i = \sum x_i y_i$$

and by (20) and (21), we have

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

$$= \sqrt{\frac{\sum x_i^2}{\sum y_i^2}} + \sqrt{\frac{\sum y_i^2}{\sum x_i^2}} - C$$

Hence

$$C\left(1 + \frac{1}{J}\right) = \sqrt{\frac{\sum x_i^2}{\sum y_i^2}} + \sqrt{\frac{\sum y_i^2}{\sum x_i^2}}$$

from which (30) readily follows. □

We now have the following basic result.

Theorem II.2:

Let $\left\|\vec{X}\right\| = a\left\|\vec{Y}\right\|$ where $a > 0$. Then

$$J = \frac{C}{a + \frac{1}{a} - C} \tag{31}$$

Proof:

This follows readily from (30) and

$$\left\|\vec{X}\right\| = \sqrt{\sum x_i^2} = a\left\|\vec{Y}\right\| = a\sqrt{\sum y_i^2} \qquad \square$$

Corrolary II.3:

If $a = 1$ (i.e. if $\left\|\vec{X}\right\| = \left\|\vec{Y}\right\|$) then

$$J = \frac{C}{2 - C} \tag{32}$$

Proof:

This follows readily from Theorem II.2. $\square$

We have calculated the values of J, obtained from (32) for increments of C equal to 0.1. The result can be seen in Table 1.

Tabel 1. Values of J versus C of formula (31) (case a=1)

C	J
0	0
0.1	0.052631
0.2	0.1111111
0.3	0.1764706
0.4	0.25
0.5	0.3333333
0.6	0.4285714
0.7	0.5384615
0.8	0.6666667
0.9	0.8181818
1	1

This function, hence, fits almost exactly Fig.1 – see Fig.2, where the curve (32) is added on the graph in Fig.1.

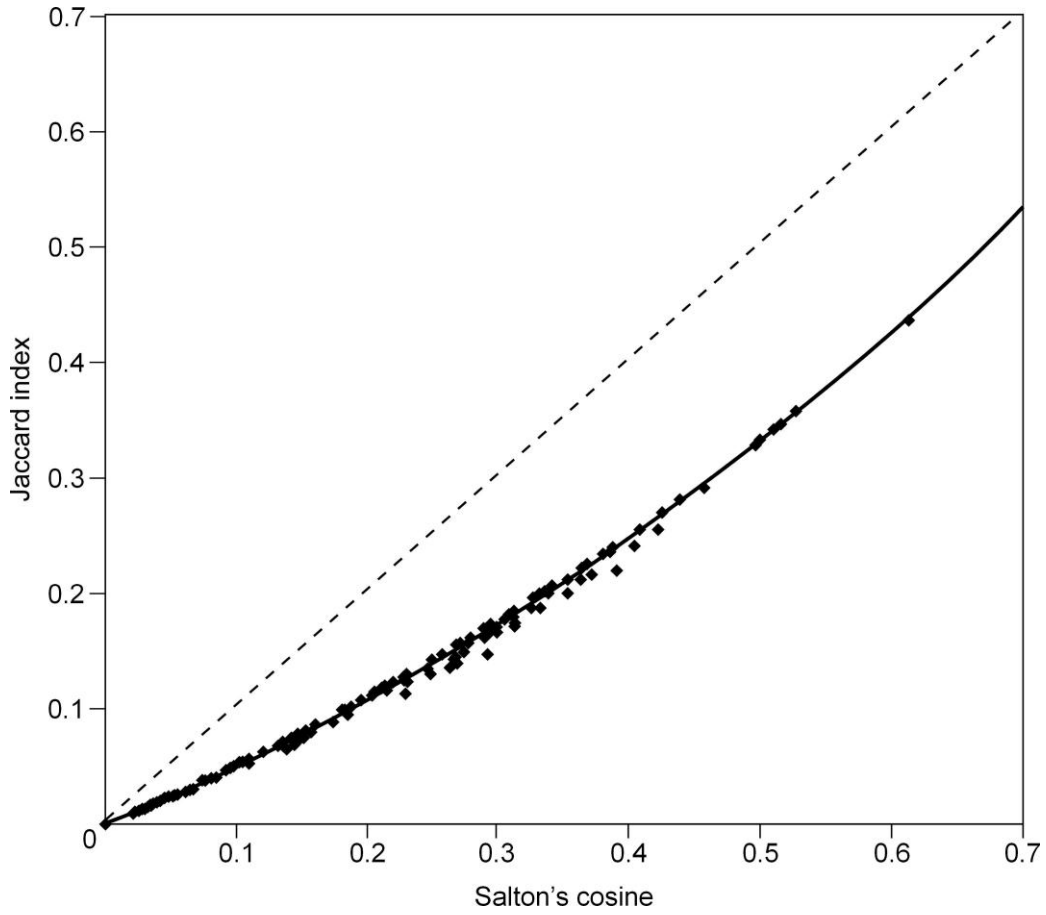

Fig. 2 Fit of (32) on the graph in Fig.1

Further to the study of the general function (31) we have the following results.

Propositon II.4:

The function (31) (hence also (32)) is a convexly increasing function of C with $J(0)=0$ and

$$J(1)=\frac{1}{a+\frac{1}{a}-1}=\frac{a}{a^2-a+1}\le 1$$

and $J(1)=1$ if and only if $a=1$.

Proof:

This follows readily from (32) and by the calculation of $\frac{dJ}{dC}$ and $\frac{d^2J}{dC^2}$ and since $\frac{a}{a^2-a+1}=1$ if and only if $(a-1)^2=1$ hence if and only if $a=1$. □

To distinguish between the functions (31) for different a we will, in the sequel, denote (31) by

$$J = J_a(C) = \frac{C}{a + \frac{1}{a} - C} \tag{33}$$

Proposition II.5:

For any $a, a' > 0$ we have

$$J_a(C) > J_{a'}(C) \tag{34}$$

for every $C \in]0,1]$ (note that $J_a(0) = J_{a'}(0) = 0$ for all $a, a' > 0$) if and only if

$$a' > \max\left(a, \frac{1}{a}\right) \tag{35}$$

or

$$a' < \min\left(a, \frac{1}{a}\right) \tag{36}$$

Furthermore, if (35) nor (36) are true, we have the opposite inequality $(\leq)$ in (34), for all $C \in [0,1]$ and $<$ in (34) for all $C \in]0,1]$, if $a \neq a'$ and $a \neq \frac{1}{a'}$.

Proof:

First note that, for all $a, a' > 0$ we have that

$$a + \frac{1}{a} - C > 0 \tag{37}$$

$$a' + \frac{1}{a'} - C > 0 \tag{38}$$

since $a + \frac{1}{a} \geq 2$ and $a' + \frac{1}{a'} \geq 2$ and since $C \in [0,1]$. Indeed

$$a + \frac{1}{a} \geq 2$$

if and only if $a^2 - 2a + 1 \geq 0$ which is true since $(a-1)^2 \geq 0$. Then (34) is equivalent with

$$a' + \frac{1}{a'} > a + \frac{1}{a} \tag{39}$$

or

$$a'a(a' - a) > a' - a \tag{40}$$

(i) Let $a' > a$. Then (40) is equivalent with $a'a > 1$ or $a' > \frac{1}{a}$. In this case $a' > \max\left(a, \frac{1}{a}\right)$.

(ii) Let $a' < a$. Then (40) is equivalent with $a'a < 1$ or $a' < \frac{1}{a}$. In this case $a' < \min\left(a, \frac{1}{a}\right)$.

Both cases together are all cases in which (34) is valid. Indeed if (35) nor (36) are valid then we have $\frac{1}{a} \le a' \le a$ or $a \le a' \le \frac{1}{a}$ in which case we see, by (40) that $J_a(C) \le J_{a'}(C)$, for all $C \in [0,1]$, the opposite of (34), with strict inequality for $C \in]0,1]$ and $a \ne a'$ and $a \ne \frac{1}{a'}$. □

Corollary II.6:

For all $C \in]0,1]$

$$J_1(C) > J_a(C) \tag{41}$$

for all $a > 0$ and $a \ne 1$.

Proof:

Since $a' = 1$ is between a and $\frac{1}{a}$, for every $a \ne 1$,we have that (35) and (36) are false, hence, by the above Theorem, we have $J_1(C) \ge J_a(C)$ for all $C \in [0,1]$, but, since $a \ne 1 = a' = \frac{1}{a'}$, we have that (41) is valid, for all $C \in]0,1]$. □

Further we can also prove the following proposition.

Proposition II.7:

For all $a > 0$ we have

$$J_a(C) \le C \tag{42}$$

for all $C \in [0,1]$, $J_a(0) = 0$,

$$J_a(1) = \frac{1}{a + \frac{1}{a} - 1} \le 1 \tag{43}$$

and where the inequality $\le$ in (43) is strict if and only if $a \ne 1$.

Proof:

By (33) we have that

$$J_a(C) \le C$$

for all $C \in [0,1]$ if and only if

$$\frac{C}{a + \frac{1}{a} - C} \leq C$$

By (37) we have then that this is equivalent with (for $C \in]0,1]$)

$$a + \frac{1}{a} - 1 \geq C \tag{44}$$

But $a + \frac{1}{a} \geq 2$ (since $(a-1)^2 \geq 0$) hence, since $C \in]0,1]$, we have that, hence (42) is true.

Hence also (43) follows and, if $a \neq 1$ the inequality (43) is strict since $(a-1)^2 > 0$. □

Note that (42) also follows from (41) and the fact that

$$J_1(C) = \frac{C}{2 - C} \leq C$$

for all $C \in [0,1]$.

Summarizing, we have that all values $J_a(0) = 0$ that only $J_1(1) = 1$ and $J_a(1) < 1$ for all $a \neq 1$ and that $J_a(C) < C$ for all $a > 0$ and $C \in]0,1[$. This supports the experimental finding in Leydesdorff (2008) that "the Jaccard index covers a "smaller range" than does the cosine" where also reference is given to Hamers, Hemeryck, Herweyers, Janssen, Keters and Rousseau (1989) stating that $C \approx 2J$ in most practical cases. This is, of course a too rough estimation but some support is given by Table 1 (case $a = 1$).

We close this section on the relation between the Jaccard measure and Salton's cosine measure by calculating the difference $J_a(C) - J_{a'}(C)$ for different values of $a, a' > 0$ and $C \in [0,1]$ (note that, always, $J_a(0) = J_{a'}(0) = 0$).

Suppose a and a' are such that $J_a(C) - J_{a'}(C) > 0$ for all $C \in]0,1]$. We have

$$J_a(C) - J_{a'}(C) = \frac{C}{a + \frac{1}{a} - C} - \frac{C}{a' + \frac{1}{a'} - C}$$

$$= \frac{C\left(a' + \frac{1}{a'} - a - \frac{1}{a}\right)}{\left(a + \frac{1}{a} - C\right)\left(a' + \frac{1}{a'} - C\right)} \tag{45}$$

in which all factors are positive, because of (37), (38) and (39).

An elementary calculation shows that (45) increases in C, implying that the largest difference occurs in $C = 1$:

$$J_a(1) - J_{a'}(1) = \frac{a' + \frac{1}{a'} - a - \frac{1}{a}}{\left(a + \frac{1}{a}\right)\left(a' + \frac{1}{a'}\right)} \tag{46}$$

We will now use (45) to estimate the range of a in $\left\|\vec{X}\right\| = a\left\|\vec{Y}\right\|$ that occurs in Leydesdorff (2008). A graphical inspection of Fig.2 gives $J_1(0.3) - J_a(0.3) \approx 0.03$ (replacing a by 1 and a' by a). Hence (45) gives (using (41))

$$J_1(0.3) - J_a(0.3) = \frac{0.3\left(a + \frac{1}{a} - 2\right)}{(2 - 0.3)\left(a + \frac{1}{a} - 0.3\right)} \leq 0.03$$

yielding

$$a + \frac{1}{a} \leq 2.3481928$$

This, in turn, yields

$$0.5588709 \leq a \leq 1.7893218$$

Note that $0.5588709 = \frac{1}{1.7893218}$ as it should. The part below $a = 1$ can be omitted since $\left\|\vec{X}\right\| = a\left\|\vec{Y}\right\|$ if and only if $\left\|\vec{Y}\right\| = \frac{1}{a}\left\|\vec{X}\right\|$. So, upon interchange of $\vec{X}$ and $\vec{Y}$ we can say that in the data in Leydesdorff (originating from Ahlgren, Jarneving and Rousseau (2003)) we have that the maximal norm factor is 1.7893218 or, roughly speaking, $\left\|\vec{X}\right\| \leq 2\left\|\vec{Y}\right\|$. This explains why a relative large variation in vector norms (up to a factor 2) still leads to very close curves $J_a(C)$ so that the cloud of points in Fig.1 is very sharp, strongly resembling a functional relation.

This ends our study of the comparison of J and C. Now we will compare all the other similarity measures with $C = Cos$.

III. The relation between the other similarity measures E, E_α, N, O_1, O_2, P and R with Cos in case $\left\|\vec{X}\right\| = a\left\|\vec{Y}\right\|$

III.1 E, E_α versus $Cos = C$

Since $E = E_{\frac{1}{2}}$, this special case will be comprised in the general one for $\alpha \in]0,1[$.

By (21) and (23) we have

$$\frac{C}{E_\alpha} = \frac{\alpha\sum x_i^2 + (1-\alpha)\sum y_i^2}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$$

$$= \alpha\frac{\sqrt{\sum x_i^2}}{\sqrt{\sum y_i^2}} + (1-\alpha)\frac{\sqrt{\sum y_i^2}}{\sqrt{\sum x_i^2}}$$

In case $\left\|\vec{X}\right\| = a\left\|\vec{Y}\right\|$ we hence have

$$\frac{C}{E_\alpha} = \alpha a + (1-\alpha)\frac{1}{a} \tag{47}$$

Hence E_α is a linear function of C

$$E_\alpha = \frac{1}{\alpha a + (1-\alpha)\frac{1}{a}}C \tag{48}$$

and for $a = 1$ this even reduces to equality: $E_\alpha = C$ for all $\alpha \in]0,1[$.

Now $\alpha a + (1-\alpha)\frac{1}{a} \geq 1$. Indeed

$$\alpha a^2 - a + (1-\alpha) = a(\alpha a - 1) - (\alpha - 1) \geq 0$$

since $\alpha a - 1 \geq \alpha - 1$ and $a \geq 1$. Hence $E_\alpha \leq C$ for all $\alpha \in]0,1[$ and $E_\alpha < C$ except if $a = 1$ ($E_\alpha = C$) and except in $C = 0$ (then $E_\alpha = 0$ also). Note that (48) reduces, for $\alpha = \frac{1}{2}$, to

$$E = \frac{2}{a + \frac{1}{a}} C \tag{49}$$

III.2 N versus E and C

It is a bit easier to, firstly, compare N with E. We have, by (22) and (24)

$$\frac{N}{E} = \frac{\sqrt{2}}{2} \frac{\sum x_i^2 + \sum y_i^2}{\sqrt{\left(\sum x_i^2\right)^2 + \left(\sum y_i^2\right)^2}}$$

Hence

$$2\left(\frac{N}{E}\right)^2 = \frac{\left(\sum x_i^2\right)^2 + \left(\sum y_i^2\right)^2 + 2\left(\sum x_i^2\right)\left(\sum y_i^2\right)}{\left(\sum x_i^2\right)^2 + \left(\sum y_i^2\right)^2}$$

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

Hence

$$2\left(\frac{N}{E}\right)^2 = \frac{a^2 + \frac{1}{a^2} + 2}{a^2 + \frac{1}{a^2}}$$

so that

$$N = \sqrt{\frac{1}{2} \frac{a^2 + \frac{1}{a^2} + 2}{a^2 + \frac{1}{a^2}}} E \tag{50}$$

which implies that $N = E$ if $a = 1$. In function of C we have, by (49), the linear function

$$N = \sqrt{\frac{2}{a^2 + \frac{1}{a^2}}} C \tag{51}$$

which, of course also implies $N = C$ if $a = 1$. Of course, since $a^2 + \frac{1}{a^2} \geq 2$ for all $a > 1$ we have that $N \leq C$.

III.3 O_1 and O_2 versus C

From (21) and (25) we have

$$\frac{C}{O_1} = \frac{\min\left(\sum x_i^2, \sum y_i^2\right)}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \tag{51}$$

In case $\left\|\vec{X}\right\| = a\left\|\vec{Y}\right\|$ we hence have

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

$$= \min\left(a, \frac{1}{a}\right)$$

hence

$$O_1 = \frac{1}{\min\left(a, \frac{1}{a}\right)} C \tag{52}$$

a linear function of C.

Similarly, using (21) and (26) we find the linear function of C

$$O_2 = \frac{1}{\max\left(a, \frac{1}{a}\right)} C \tag{53}$$

which readily implies that, always, $O_1 \geq C$ and $O_2 \leq C$ but also that, if $a = 1$, that $O_1 = O_2 = C$. Note that, if $a \neq 1$, $\max\left(a, \frac{1}{a}\right) > 1$ and $\min\left(a, \frac{1}{a}\right) < 1$ hence $O_1 > C$ and $O_2 < C$ for all $a > 0$, $a \neq 1$. This also goes for $C = 1$ in which case $O_2 < 1$ but $O_1 > 1$. Note that this is only possible in the non-binary case. If $C = 1$ then we have that $\vec{X} = a\vec{Y}$ for a certain $a \neq 0$ (since $C = 1$ implies that the angle between $\vec{X}$ and $\vec{Y}$ is zero). But, in the binary case, $\vec{X} = a\vec{Y}$ implies $a = 1$ and hence $O_1 = C = 1$.

III.4 P and R versus C

(21), (27) and (28) imply

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

$$\frac{C}{P} = \frac{\sqrt{\sum x_i^2}}{\sqrt{\sum y_i^2}} \tag{54}$$

and

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

$$\frac{C}{R} = \frac{\sqrt{\sum y_i^2}}{\sqrt{\sum x_i^2}} \tag{55}$$

In the case that $\left\|\vec{X}\right\| = a\left\|\vec{Y}\right\|$ we hence have the linear functions

$$P = \frac{1}{a}C \tag{56}$$

$$R = aC \tag{57}$$

hence $P \geq C$ if and only if $R \leq C$. Again, for $a = 1$ we have $P = R = C$.

Remark: It is clear that from the proved relations between any measure and C we can also calculate all other relations between any two measures. It is clear that, in case $\left\|\vec{X}\right\| = a\left\|\vec{Y}\right\|$ ($a > 0$ a constant), from the linear relations with C we will also obtain linear relations between any two measures (excluding J).

IV. Conclusions and suggestions for further research

We defined the similarity measures J (Jaccard), $\mathrm{Cos} = \mathrm{C}$ (Salton's Cosine), E and E_α ((generalized) Dice), N and the overlap measures O_1, O_2 (symmetric) and P and R (non-symmetric).

On the "trajectories" $\left\|\vec{\mathrm{X}}\right\| = \mathrm{a}\left\|\vec{\mathrm{Y}}\right\|$ with $\mathrm{a} > 0$ constant we studied the relations among these similarity measures and we showed that J is a convexly increasing function of C, hereby presenting a model that explains Fig.1, given in Leydesdorff (2008). We even explain that variations of a (up to a factor 2) lead to very small changes in the relation between J and C (explaining the sharp function in Leydesdorff (2008)).

On the same trajectories we show that all other measures are a linear function of C and we can even prove that they are all equal to C in case $\mathrm{a} = 1$.

In the literature one sometimes finds other similarity measures, based on $\sum_{i=1}^{n} \mathrm{x}_i$ and $\sum_{i=1}^{n} \mathrm{y}_i$, instead of $\sum_{i=1}^{n} \mathrm{x}_i^2$ and $\sum_{i=1}^{n} \mathrm{y}_i^2$. So can one find in Jones and Furnas (1987) the "pseudo" cosine and Dice measures

$$\mathrm{PCos} = \frac{\sum_{i=1}^{n} \mathrm{x}_i \mathrm{y}_i}{\left(\sum_{i=1}^{n} \mathrm{x}_i\right)\left(\sum_{i=1}^{n} \mathrm{y}_i\right)} \tag{58}$$

and

$$\mathrm{PE} = \frac{\sum_{i=1}^{n} \mathrm{x}_i \mathrm{y}_i}{\sum_{i=1}^{n} \mathrm{x}_i + \sum_{i=1}^{n} \mathrm{y}_i} \tag{59}$$

, i.e. using the L^1-norms $\left\|\vec{X}\right\|_1 = \sum_{i=1}^{n} x_i$, $\left\|\vec{Y}\right\|_1 = \sum_{i=1}^{n} y_i$ (for x_i , $y_i \geq 0$) instead of the (in this article) used L^2-(or Euclidean) norms $\left\|\vec{X}\right\| = \left\|\vec{X}\right\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$ and $\left\|\vec{Y}\right\| = \left\|\vec{Y}\right\|_2 = \sqrt{\sum_{i=1}^{n} y_i^2}$. The norms $\|\cdot\|_1$ correspond to the so-called "city-block" metric (see e.g. Egghe and Rousseau (1990), Chapter I, the section on multivariate statistics).

Similarly, one finds in Dominich (2001) the $\|\cdot\|_1$ version of O_1:

$$PO_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\min\left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i\right)} \tag{60}$$

(while the similar variant of O_2 is not discussed in Dominich (2001)).

One can even think of defining <u>all</u> the other measures, used here, in this way. Note that such definitions equally extend the binary case in the same way. Indeed (17) and (18) can be replaced by

$$|A| = \sum_{i=1}^{n} x_i = \left\|\vec{X}\right\|_1 \tag{61}$$

$$|B| = \sum_{i=1}^{n} y_i = \left\|\vec{Y}\right\|_1 \tag{62}$$

as well.

If we use the trajectories $\left\|\vec{X}\right\|_1 = a\left\|\vec{Y}\right\|_1$ now, then all the results obtained in this paper can be reproved for the "pseudo" versions of these measures (such as (58), (59) or (60)). Note that the term pseudo refers to the fact that they are variants of the original definition but that these measures are good similarity measures, as well.

We do not see, at the moment, how to relate the $\|\cdot\|_2$ -defined measures with the $\|\cdot\|_1$ -defined ones. Also, we do not see how the classical correlation coefficient of Pearson, fits in this theory. This measure is defined as (cf. Egghe and Rousseau (1990), Ahlgren, Jarneving and Rousseau (2003)).

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \qquad (63)$$

The problem to relate r with other measures is the simultaneous occurrence of $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$ (and similarly for $\vec{Y}$). In Ahlgren, Jarneving and Rousseau (2003) it is also shown that r lacks some properties that good similarity measures should have. Due to the lack of the inproduct $\sum_{i=1}^{n} x_i y_i$ in the χ^2-measure, defined in Ahlgren, Jarneving and Rousseau (2003)

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{x_i}{\sum_{j=1}^{n} x_j} - \frac{y_i}{\sum_{j=1}^{n} y_j}\right)^2 \qquad (64)$$

we also leave it open to relate χ^2 with the other measures.

We want to conclude that from the definitions of the diverse measures, formal mathematical relationships follow and therefore should not be considered as sociological phenomena. In other words, the importance of studying different similarity measures decreases and drawing sociological conclusions on these relations does not make much sense. It also follows that one measure is enough for ranking.

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. New York, Academic Press.

S. Dominich (2001). Mathematical Foundations of Information Retrieval. Kluwer Academic Publishers, Dordrecht.

L. Egghe (2007). Existence theorem of the quadruple (P,R,F,M): Precision, Recall, Fallout and Miss. Information Processing and Management 43(1), 265-272.

L. Egghe (2008). The measures recall, precision, fallout and miss in function of the number of retrieved documents and their mutual interrelations. Information Processing and Management, to appear.

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.

D.A. Grossman and O. Frieder (1998). Information Retrieval Algorithms and Heuristics. Kluwer Academic Publishers, Boston.

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

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

L. Leydesdorff (2008). 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.

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

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

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

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