Brief Communication

On the relation between the Association Strength and other similarity measures

Leo 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

A graph in van Eck and Waltman [JASIST 60(8), 2009, p. 1644], representing the relation between the Association Strength and the Cosine, is partially explained as a sheaf of parabolas, each parabola being the functional relation between these similarity measures on the trajectories $\vec{X}.\vec{Y} = a$, a constant.

Based on earlier obtained relations between Cosine and other similarity measures (such as Jaccard index) we can prove new relations between the Association Strength and these other measures.

---

[1] Permanent address

Key words and phrases: association strength, Cosine, Jaccard, inclusion, overlap
Acknowledgement: The author is grateful to Nees Jan van Eck and Ludo Waltman for interesting discussions on the topic of this paper.

I. Introduction

Occurrence matrices describe the occurrence of objects (e.g. keywords) in e.g. documents. Documents are rows and objects are columns. The value o_{ij} on the coordinate (i, j) hence gives the number of times object j occurs in document i (and the value is zero if object j does not occur in document i). Instead of the actual value o_{ij} of number of occurrences of object j in document i one can give the value $o_{ij} = 1$ if object j occurs in document i (how many times is not important here). Then we have a so-called binary occurrence matrix. This will not be supposed in this paper: o_{ij} denotes the number of occurrences of object j in document i. Other example: rows are web-pages and columns are hyperlinks (or, again, key words). For more examples, see van Eck and Waltman (2009).

In occurrence matrices one can calculate similarities between each pair of columns, being vectors $\vec{X} = (x_1, ..., x_N)$, $\vec{Y} = (y_1, ..., y_N)$. This can be done using diverse similarity measures. For an overview, see Egghe (2009). Also in Egghe (2009) one studied relations between these similarity measures based on remarkable shapes of clouds of points (each point showing the values of two of these similarity measures for a couple of vectors $\vec{X}, \vec{Y}$).

In Egghe (2009) the main result was on the relation between the Jaccard index J and the Cosine C defined as

$$J = \frac{\vec{X}.\vec{Y}}{\| \vec{X} \|_2^2 + \| \vec{Y} \|_2^2 - \vec{X}.\vec{Y}} \tag{1}$$

and

$$C = \frac{\vec{X}.\vec{Y}}{\| \vec{X} \|_2 \| \vec{Y} \|_2} \tag{2}$$

where $\vec{X}$ and $\vec{Y}$ are as above and where

$$\vec{X}.\vec{Y} = \sum_{i=1}^{N} x_i y_i \tag{3}$$

is the inner product of $\vec{X}$ and $\vec{Y}$ and where

$$\| \vec{X} \|_2^2 = \sum_{i=1}^{N} x_i^2 \tag{4}$$

$$\| \vec{Y} \|_2^2 = \sum_{i=1}^{N} y_i^2 \tag{5}$$

are the square of the L^2-norms of $\vec{X}$ and $\vec{Y}$. Henceforth we will simplify the notation $\sum_{i=1}^{N}$ into $\sum$. So we have

$$J = \frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i y_i} \tag{6}$$

$$C = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \tag{7}$$

A functional relation between J and C does not exist. However, in Egghe (2009) we could prove that, on the trajectories $\| \vec{X} \|_2 = a \| \vec{Y} \|_2$ (a is a fixed parameter), there is a functional relation between J and C, namely (see Theorem II.2, p. 235)

$$J = \frac{C}{a + \frac{1}{a} - C} \tag{8}$$

which shape (convexly increasing) was confirmed in the graphs in Leydesdorff (2008).

A similar argument also showed that, on the trajectories $\| \vec{X} \|_2 = a \| \vec{Y} \|_2$ we have that

$$O_1 = \frac{1}{\min\left(a, \frac{1}{a}\right)} C \tag{9}$$

where

$$O_1 = \frac{\sum x_i y_i}{\min\left(\sum x_i^2, \sum y_i^2\right)} \tag{10}$$

which is an overlap measure (see Egghe (2009)) and is called the Inclusion index in van Eck and Waltman (2009).

Formula (9) is a linear relationship between O_1 and C. For other relations, see Egghe (2009).

In van Eck and Waltman (2009) one studies the measures C, J and O_1 and one more measure (not studied in Egghe (2009)): the so-called Association strength (denoted here by A).

$$A = \frac{\sum x_i y_i}{\sum x_i^2 \sum y_i^2} \tag{11}$$

It is, however, not possible to find a functional relation between A and the other measures using the trajectories $\| \vec{X} \|_2 = a \| \vec{Y} \|_2$.

In the next section another "trick" will establish a functional relationship between A and the other three measures which are studied here. This will confirm some graphical relations between A and the other measures as found in van Eck and Waltman (2009).

II. Relations of A versus C, O_1 and J

When we look at Fig. 1, a graph in van Eck and Waltman (2009), it is clear that the relation between A and C is a collection of parabolas of varying width. This indicates a quadratic

"relation" of A (ordinate) versus C (abscissa), although a functional relation between A and C (in fact between any two similarity measures) does not exist.

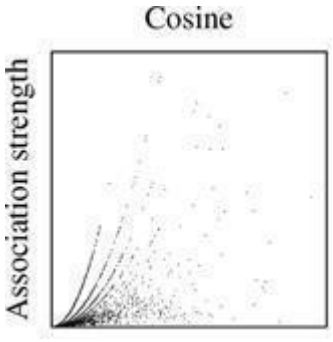

Fig. 1 The relation between the Association Strength A and the Cosine C: a sheave of parabolas.
Reproduced with kind permission of Wiley-Blackwell

Therefore we will use the square of C:

$$C^2 = \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2 \sum y_i^2} \tag{12}$$

So, from (11) and (12) we see that

$$\frac{C^2}{A} = \sum x_i y_i \tag{13}$$

We now suppose that we are on the trajectory $\vec{X}.\vec{Y} = b$ (b a fixed parameter). We now have

$$A = \frac{1}{b} C^2 \tag{14}$$

which are indeed convex parabolas with minimum through $(0,0)$ and with varying width, expressed by the variable parameter $\frac{1}{b}$ (we mean: in (14), b is fixed but we let then b vary so that we obtain a sheaf of parabolas).

In Fig. 1 we see that the "distance" between the parabolas is not constant. Also this can be explained. In van Eck and Waltman (2009) the vectors $\vec{X}$ and $\vec{Y}$ are binary: they only have coordinates 0 or 1. Since $b = \vec{X}.\vec{Y}$ this means that $b \in \{1, 2, ..., N\}$, where N is the dimension of the vectors. Hence

$$\frac{1}{b} \in \left\{ \frac{1}{N}, \frac{1}{N-1}, ..., \frac{1}{2}, 1 \right\}$$

explaining the variable "distances" between the parabolas in Fig. 1

This explains the $A(C)$ relationship.

For the relationship between A and O_1 (Inclusion index) we use (9) (under the condition $\| \vec{X} \|_2 = a \| \vec{Y} \|_2$), together with (14) (under the condition $\vec{X}.\vec{Y} = b$). Note that the two conditions together still leave an infinite number of cases of vectors $\vec{X}$ and $\vec{Y}$. We now have

$$A = \frac{\min\left(a^2, \frac{1}{a^2} \right)}{b} O_1^2 \tag{15}$$

The graph in van Eck and Waltman (2009) on the $A(O_1)$ relation is more diffuse than the $A(C)$ relation probably due to the occurrence of the two parameters a and b.

Finally we look for the $A(J)$ relation. From (8) we have

$$C = \frac{a + \frac{1}{a}}{1 + \frac{1}{J}} \tag{16}$$

Hence (16) in (14) yields

$$A = \frac{\left(a + \frac{1}{a}\right)^2}{b} \frac{J^2}{\left(1+J\right)^2} \tag{17}$$

So, this is (for variable a and b) a sheaf of functions which have the form of

$$f\left(x\right) = \frac{x^2}{\left(1+x\right)^2} \tag{18}$$

It is readily seen that

$$f'\left(x\right) = \frac{2x + 2x^2}{\left(1+x\right)^4} > 0$$

$$f''\left(x\right) = \frac{\left(1+x\right)^4\left(2+4x\right) - \left(2x+x^2\right)4\left(1+x\right)^3}{\left(1+x\right)^8}$$

which has the sign of the numerator which equals

$$-4x^5 - 14x^4 - 16x^3 - 4x^2 + 4x + 2$$

This polynomial is zero in $x = \frac{1}{2}$, positive for $0 < x < \frac{1}{2}$ and negative for $\frac{1}{2} < x < 1$.

So (17) represents a sheaf of functions which are strictly increasing, first convex (positive second derivative), then concave (negative second derivative). In the graph in van Eck and Waltman (2009) (given here in Fig. 2) again (due to the occurrence of two parameters, probably) the cloud of points is rather confuse but we clearly see the initial convex part.

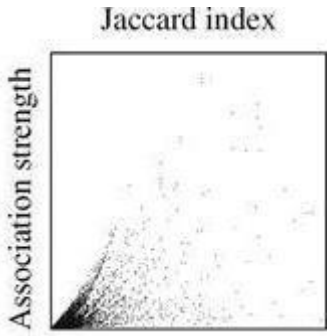

Fig. 2 The relation between the Association Strength A and Jaccard's index. Reproduced with kind permission of Wiley-Blackwell

Note that the in Egghe (2009) proved relation (8) between J and C is also refound in van Eck and Waltman (2009): a convexly increasing relation. Also the linear relation (9) between O_1 and C is clearly seen in van Eck and Waltman (2009).

For a study of the relation between C and Pearson's correlation coefficient, we refer the reader to Egghe and Leydesdorff (2009)).

References

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 L. Leydesdorff (2009). The relation between Pearson's correlation coefficient r and Salton's cosine measure. Journal of the American Society for Information Science and Technology 60(5), 1027-1036.

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.