Classical retrieval and overlap measures such as Jaccard's coefficient, Salton's cosine measure and the Dice coefficient satisfy the requirements for rankings based upon a Lorenz curve.

Classical retrieval and overlap measures such as Jaccard's coefficient, Salton's cosine measure and the Dice coefficient satisfy the requirements for rankings based upon a Lorenz curve.

Leo Egghe

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium
& UA, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
Email: leo.egghe@luc.ac.be.

Ronald Rousseau

KHBO, IWT, Zeedijk 101, B-8400 Oostende, Belgium
& UA, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
Email: ronald.rousseau@khbo.be

Abstract

Classical information retrieval and overlap measures such as the Jaccard index, the Dice coefficient and Salton's cosine measure can be characterized by Lorenz curves. This result demonstrates the existence of a link between information retrieval and the information sciences on the one hand, and concentration and diversity theory, as used, e.g., in social economics and ecology on the other.

Keywords: information retrieval, overlap studies, presence-absence data, Jaccard coefficient, Salton's cosine measure, Dice coefficient, Lorenz curves, Gini index

Introduction

The use of presence-absence data is a basic approach for representing categorical data. As presence is often represented by 1 and absence by 0, such data are also known as 1 - 0 data. Their binary nature is especially suited for computing, and makes their use in science ubiquitous. This is also the case in the information sciences (Huot et al., 1992; Magurran, 1991). Indeed, a standard approach to information retrieval and the study of overlap represents documents by an array of keywords and/or phrases. Note that we chose the term 'array' and not 'vector' because these entities are not vectors in the strict mathematical sense of the word. If keywords are chosen in a particular order, a document is represented by a presence-absence (1-0) array. Similarity between documents is determined by comparing these document representations. In information retrieval and in overlap studies it is customary not to consider common zeros (Salton & McGill, 1983). Indeed, keywords or phrases that do not occur in at least one of the documents do not make the documents more similar. Medical articles are not more similar if they both do not use the keyword "Tanzania". For this reason we will refer to this approach as the zero insensitive case. This will be the only case studied in this article.

Another way of representing such data is through set theory. An entity such as a document is represented by the set of properties it possesses (here keywords or phrases present in the document, or by which the document is indexed). Note that absence data are usually not explicitly represented in the set-theoretic

approach. This corresponds to the fact that in the array representation, common zeros are not considered. Figure 1 illustrates these different approaches.

DOCUMENT A
Patents
Citations
Mapping

DOCUMENT B
Citations
Publications

Fig.1a : Two document representations (using keywords)

DOCUMENT A	X	X	X	
DOCUMENT B		X		X

=

DOCUMENT A	1	1	1	0
DOCUMENT B	0	1	0	1

Fig.1b: The same documents represented as presence-absence data (in the order: patents – citations – mapping – publications)

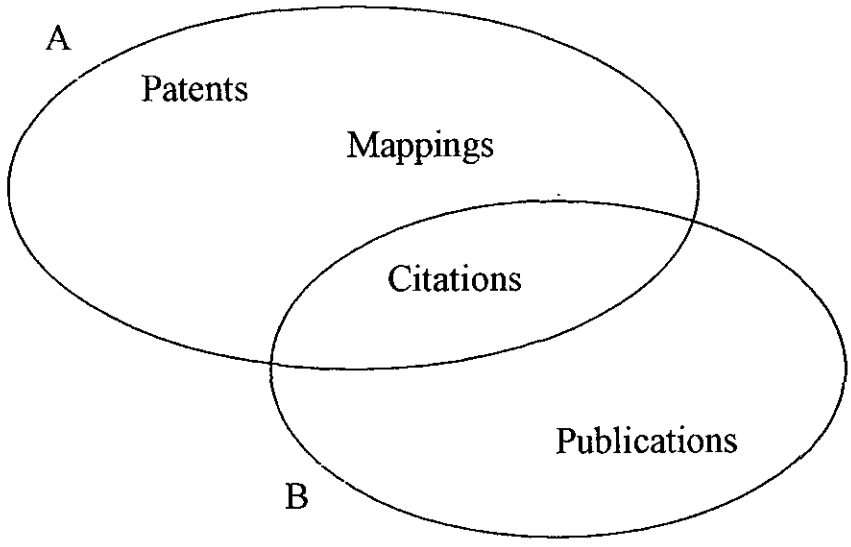

Fig. 1c The same documents represented as sets

We will show how a Lorenz curve approach leads to an intrinsic notion of similarity. These Lorenz similarity curves define a partial order in the set of presence-absence data representations. We will further show that for zero insensitive presence-absence data well-known measures such as the Jaccard index, the Dice coefficient and Salton's cosine measure respect this partial order, hereby revealing new good properties of these important measures.

Lorenz similarity curves

Let $r = (x_i)_{i=1,\ldots,N}$ and $s = (y_i)_{i=1,\ldots,N}$ be two presence-absence arrays of length N (in short: N-arrays) for the documents r and s. The similarity of $D = \{r,s\}$ must not depend on the order in which we consider r and s. It must, moreover, not depend on the order in which the keywords of r and s are enumerated. Of course, in practice one uses a particular order (always the same for the two items involved) but the point is that this must not influence their similarity. These requirements are also imposed in concentration studies, hence it seems natural to consider Lorenz curves (Lorenz, 1905) in a similar vain as for 'concentration' and 'diversity' studies. In a first step we will construct a Lorenz curve suited to the study of similarity. Only in a second step will similarity be measured by a function respecting the partial order introduced in the first step. An infinite number of such functions are mathematically possible. In order to emphasize the fact that it is irrelevant in which order document representations for similarity studies are considered we refer to $D = \{r,s\}$ as a duo, a word that has no "rank" connotations.

Construction of Lorenz curves for duo similarity

Let $r = (x_i)_{i=1,\ldots,N}$ and $s = (y_i)_{i=1,\ldots,N}$ be two presence-absence N-arrays of the documents r and s (such arrays will also be called keyword arrays) and let $(a_i)_{i=1,\ldots,N}$ and $(b_i)_{i=1,\ldots,N}$ denote their relative arrays. This means that each value is divided by the total sum, and hence the new sum of all components becomes equal to one. Formally:

$$\forall i = 1,\ldots,N : a_i = \frac{x_i}{T_r},\ b_i = \frac{y_i}{T_s} \tag{1}$$

with T_r equal to $\sum_{j=1}^{N} x_j$ (the total sum of the r-array) and T_s equal to $\sum_{j=1}^{N} y_j$ (the total sum of the s-array). Next, the components of the difference array $d = (d_i)_{i=1,\ldots,N}$ with $d_i = a_i - b_i$ are ranked from largest to smallest.

Finally, putting

$$c_i = \sum_{j=1}^{i} d_j = \sum_{j=1}^{i} \left(a_j - b_j\right) \tag{2}$$

the Lorenz similarity curves are obtained by joining the origin (0,0) with the points with coordinates

$$\left(\frac{i}{N},\ c_i\right)_{i=1,\ldots,N} \tag{3}$$

This construction was introduced in (Egghe & Rousseau, 2001) in the more general context of so-called 'symmetric relative concentration theory'. The c-arrays will be called the (Lorenz curve) ordinate arrays. The first coordinate of equation (3) will be referred to as the x-coordinate, or the abscissa, while the second one will be called the y-coordinate or the ordinate. Note that these curves always end in the point E = (1,0) as

$$c_N = \sum_{j=1}^{N}\left(a_j - b_j\right) = \sum_{j=1}^{N} a_j - \sum_{j=1}^{N} b_j = 1 - 1 = 0 \qquad (4)$$

In this sense these Lorenz curves differ from the classical ones that end in the point (1,1). Figure 1 gives some examples of Lorenz similarity curves.

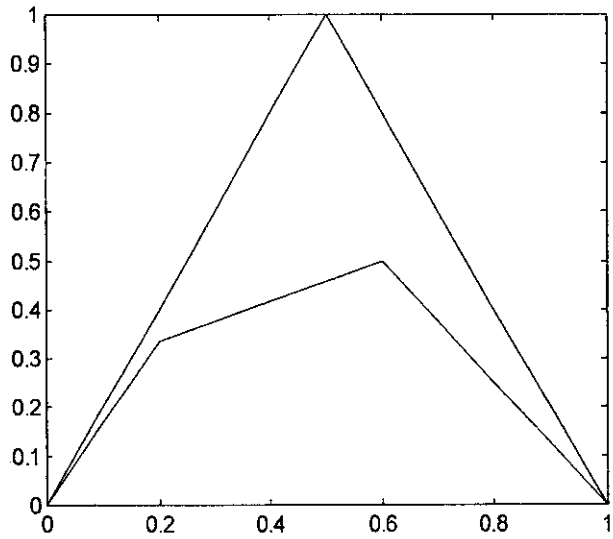

Fig.1 Lorenz curves of the duos D' = {(1,0),(0,1)}, above, and of D = {(1,1,1,1,1,1,0,0,0,0),(0,0,1,1,1,1,1,1,1,1)}, under.

Similar to other (namely the standard) Lorenz curves also these lead to partial orders and functions respecting the partial order are the ones we are interested in. Yet, these partial orders are not completely trivial. Indeed, there are two complicating factors, the first one being related to the question: what happens if we use the array d' with components $(b_j\text{-}a_j)_j$ instead of d with components $(a_i\text{-}b_i)_i$? Recall that this must lead to the same similarity. Yet, the corresponding Lorenz similarity curves are clearly different. Luckily, there exists a simple relation between these two Lorenz similarity curves: they are symmetric with respect to

the line x = 0.5, see Fig.2. (For a proof in the general case, i.e., not necessarily presence-absence data, we refer the reader to (Egghe & Rousseau, 2001)).

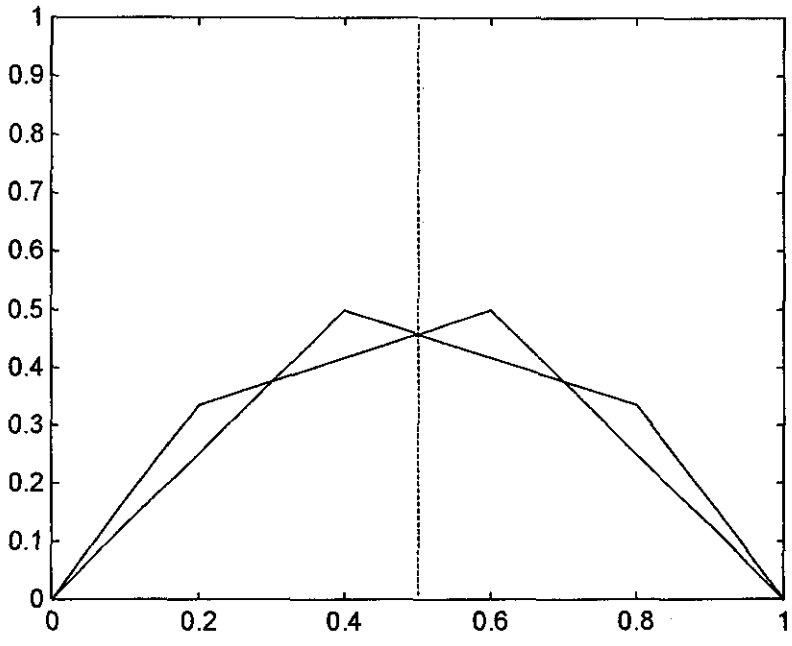

Figure 2. The similarity Lorenz curves of D = {(1,1,1,1,1,1,0,0,0,0),(0,0,1,1,1,1,1,1,1,1)} and D' = {(0,0,1,1,1,1,1,1,1,1), (1,1,1,1,1,1,0,0,0,0)} are each other's mirror image with respect to the line x = 0.5.

The second point is that many duos lead to the same Lorenz similarity curve. We will hence consider equivalence classes. Two duos are equivalent if they lead to the same Lorenz similarity curve or to curves which are reflections with respect to the line x = 1/2. The duos {r" = (1,1,1), r" = (1,1,1)} and {r"' = (1,1,1,1,1), r"" = (1,1,1,1,1)} are equivalent. This example leads to the line connecting the origin O = (0,0) and the point E = (1,0). This line will be referred to as the equality line. Other examples of equivalent duos will be given when the notion of replication invariance is introduced.

Notation

The similarity Lorenz curve of (r,s) – in this order - will be denoted by $L_{r,s}$. Consequently the Lorenz curve of (s,r) is denoted as $L_{s,r}$. As $L_{s,r}$ is the reflection of $L_{r,s}$ with respect to the line x = 0.5, $L_{s,r}$ is also denoted as $R(L_{r,s})$. Here, the symbol R stands for the reflection operation. In what follows these curves will be considered as being equivalent, and, if no distinction is necessary, they will be denoted as L_D, with D = {r,s}. A Lorenz similarity curve can be considered as a function of the abscissa. In such case it will be denoted as $L_{r,s}(x)$ or $L_D(x)$.

The partial order derived from Lorenz similarity curves

For D = {r,s} and D'= {r',s'} we will say that the Lorenz similarity curve $L_D(x)$ is situated below the Lorenz similarity curve $L_{D'}(x)$ if for every $x \in [0,1]$ either $L_{r,s}(x) \leq L_{r',s'}(x)$, with strict inequality in at least one point (and hence in infinitely many), or $L_{s,r}(x) \leq L_{r',s'}(x)$, with strict inequality in at least one point. Note that by the previous observation $L_{r,s}(x) \leq L_{r',s'}(x)$ automatically implies that $L_{s,r}(x) \leq L_{s',r'}(x)$. Similarly, we say that the Lorenz similarity curve $L_D(x)$ coincides with the Lorenz similarity curve $L_{D'}(x)$ if for every $x \in [0,1]$ either $L_{r,s}(x) = L_{r',s'}(x)$, or $L_{s,r}(x) = L_{r',s'}(x)$.

We are now in a position to define a partial order derived from Lorenz similarity curves.

Definition: Lorenz similarity

Let D = {r,s} and D' = {r',s'}. Then the Lorenz similarity between r and s is intrinsically larger than the Lorenz similarity between r' and s' if the Lorenz similarity curve of D, L_D, is situated below the Lorenz similarity curve of D', $L_{D'}$. Fig. 1 gives an illustration of such a situation.

Notation

If D = {r,s}, D' = {r',s'} and the Lorenz similarity between r and s is intrinsically larger than the Lorenz similarity between r' and s', or if the Lorenz curves coincide then we denote this fact as:

$$D \geq D' \quad \text{or} \quad \{r,s\} \geq \{r',s'\} \quad \text{or equivalently:} \quad L_{r,s} \geq L_{r',s'}$$

If the Lorenz similarity between r and s is intrinsically (strictly) larger than the Lorenz similarity between r' and s', this fact is denoted using the symbol >.

Comments

1. In (Egghe & Rousseau, 2001) the larger curve was the one situated on top. Here, the larger is the lower one, because we consider the notion of 'similarity' as a kind of opposite to the notion of 'concentration'.
2. In general, we do not require the length of r (= the length of s) to be equal to the length of r' (= length of s').
3. Lorenz similarity curves may intersect, and will often do so. Then the corresponding similarities are intrinsically incomparable (in our framework). Of course once acceptable similarity functions are defined (see further) Lorenz curves are mapped to numbers, and hence the corresponding duos usually become comparable. Within our framework intrinsic comparability based on

Lorenz curve is the fundamental notion. Comparability based on a function such as the cosine measure leads only to a second order type of comparison.

General properties of Lorenz similarity

In this section we will prove properties that are intrinsically true for Lorenz similarity. This means that these properties always hold whatever the similarity function one uses. Such properties are the most important, basic, ones for the notion of Lorenz similarity.

Theorem. Lorenz similarity is replication invariant.

Replication means that every absence-presence value is transformed to f times this value, with f a natural number larger than one. Some examples: the duo D = {(1,0,1),(1,1,0)} is transformed to D'= {(1,1,0,0,1,1),(1,1,1,1,0,0)} (for f = 2), or to D" = {(1,1,1,0,0,0,1,1,1), (1,1,1,1,1,1,0,0,0)} (for f = 3), and so on. The proof, an immediate consequence of the way in which Lorenz curves are constructed, is omitted.

As replicated arrays lead to the same Lorenz similarity curves, they are considered to be equivalent. This property shows that if the pattern of zeros and ones is not changed then the length of corresponding arrays has no influence on Lorenz similarity.

Notation (zero-insensitive case for absence-presence data)

Consider a duo D = {r,s}, with respectively |r| and |s| (both $\neq 0$) ones, where the number of ones in a general array g is denoted by |g|. Assume that these arrays

are of length N, having c ones in common. Then N = |r| + |s| - c and, considering array r first, and then s, their difference array consists of (|r| - c) times the value $|r|^{-1}$, c times the value ($|r|^{-1}$ - $|s|^{-1}$), and (|s| - c) times the value - $|s|^{-1}$.

The corresponding Lorenz similarity curve consists of three line segments (in general), connecting the points with coordinates (0,0), $\left(\frac{|r|-c}{N},\frac{|r|-c}{|r|}\right)$, $\left(\frac{|r|}{N},\frac{|s|-c}{|s|}\right)$, (1,0) (some of these may coincide, leading to two or even one line segment). Note that these are the coordinates if we consider r first and then s (otherwise the Lorenz similarity curve connects (0,0), $\left(\frac{|s|-c}{N},\frac{|s|-c}{|s|}\right)$, $\left(\frac{|s|}{N},\frac{|r|-c}{|r|}\right)$ and (1,0) in this order). With this notation it is easy to see that one curve is symmetric to the other with respect to the line x = 0.5. If |s| > |r| the top of the Lorenz curve is situated at the point with ordinate $\frac{|s|-c}{|s|}$, otherwise the top is at the point with ordinate $\frac{|r|-c}{|r|}$. We will further denote max (|r|,|s|) by σ and min (|r|,|s|) by ρ. Then, the top, denoted as T, is situated at the point with ordinate $1-\frac{c}{\sigma}$. The other non-end point of the Lorenz similarity curve will be called the sub-top and will be denoted by S. Its ordinate is always equal to $1-\frac{c}{\rho}$. If c = 0 then S coincides with T = $\left(\frac{\rho}{N},1\right)$; if c = ρ, then S coincides with O, and T =

$\left(\frac{\rho}{N}, \frac{\sigma-\rho}{\sigma}\right)$. If $\rho = \sigma$ ($\neq c$) then the ordinates of top and sub-top are equal, namely $\frac{\sigma-c}{\sigma}$ and the Lorenz similarity curve has a horizontal line segment. Finally, if $c = \rho = \sigma$, then S = O and T = E, and we obtain the equality line.

In order to simplify calculations we will normally, i.e. if we do not state otherwise, assume that Lorenz similarity curves are drawn in such a way that the abscissa of the sub-top is never strictly larger than the abscissa of the top. We will simply say that "the sub-top comes before the top". Under these assumptions the coordinates of sub-top and top are:

$$S = \left(\frac{\rho-c}{N}, \frac{\rho-c}{\rho}\right) \tag{5}$$

$$T = \left(\frac{\rho}{N}, \frac{\sigma-c}{\sigma}\right) \tag{6}$$

with $N = \sigma + \rho - c$.

With this notation we are able to prove the following important proposition.

Proposition

If $L_{r,s} > R(L_{r',s'})$, then $L_{r,s} > L_{r',s'}$

Recall that $L_{r,s}$ and $L_{r',s'}$ are drawn with the sub-top before the top.

Proof. Sub-tops and tops of $L_{r,s}$, $L_{r',s'}$ and $R(L_{r',s'})$ are denoted respectively by S, T, S', T ', S" and T". Their coordinates are:

$$S=\left(\frac{\rho-c}{N},\frac{\rho-c}{\rho}\right)(5)\ \&\ T=\left(\frac{\rho}{N},\frac{\sigma-c}{\sigma}\right)(6)\ \text{with } N=\sigma+\rho-c$$

$$S'=\left(\frac{\rho'-c'}{N'},\frac{\rho'-c'}{\rho'}\right)(7)\ \&\ T'=\left(\frac{\rho'}{N'},\frac{\sigma'-c'}{\sigma'}\right)(8)\ \text{with } N'=\sigma'+\rho'-c'$$

$$S''=\left(\frac{\sigma'}{N'},\frac{\rho'-c'}{\rho'}\right)(9)\ \&\ T''=\left(\frac{\sigma'-c'}{N'},\frac{\sigma'-c'}{\sigma'}\right)(10)\ \text{with } N'=\sigma'+\rho'-c'$$

The proof is divided into several steps.

Step 1. abscissa(S') ≤ abscissa(S)

From the fact that $L_{r,s} > R(L_{r',s'})$ it follows that the absolute value of the slope of TE is larger than or equal to the absolute value of the slope of S"E. This means that

$$\frac{\frac{\sigma-c}{\sigma}}{1-\frac{\rho}{\sigma+\rho-c}}\geq\frac{\frac{\rho'-c'}{\rho'}}{1-\frac{\sigma'}{\sigma'+\rho'-c'}}$$

$$\Leftrightarrow\ \frac{N(\sigma-c)}{\sigma(\sigma-c)}\geq\frac{N'(\rho'-c')}{\rho'(\rho'-c')}$$

$$\Leftrightarrow\ \frac{N}{\sigma}\geq\frac{N'}{\rho'}\qquad \text{if } \sigma\neq c \text{ and } \rho'\neq c'$$

As $\rho\leq\sigma$ and $\rho'\leq\sigma'$ we conclude that

$$\frac{N'}{\sigma'}\leq\frac{N'}{\rho'}\leq\frac{N}{\sigma}\leq\frac{N}{\rho}\qquad(11)$$

If $\sigma = c$ then $c = \rho = \sigma$ and $L_{r,s}$ is the equality line, which is in contradiction with $L_{r,s} > R(L_{r',s'})$. Hence this case cannot occur. The case $c' = \rho'$ will be studied in step 5 of the proof.

Now, abscissa(S') ≤ abscissa(S) holds if:

$$\frac{\rho'-c'}{N'} \leq \frac{\rho-c}{N} \Leftrightarrow \frac{N'-\sigma'}{N'} \leq \frac{N-\sigma}{N}$$
$$\Leftrightarrow \frac{\sigma'}{N'} \geq \frac{\sigma}{N} \Leftrightarrow \frac{N'}{\sigma'} \leq \frac{N}{\sigma}$$

The last inequality is true by (11). This proves this first step.

Step 2. The slope of the line OS' is smaller than or equal to the slope of OS.

The slope of the line OS' is equal to $\frac{N'}{\rho'}$, while the slope of OS is $\frac{N}{\rho}$. The required inequality again follows from (11).

Step 1 and 2 put together show that the point S' is situated below or on the line segment OS.

Step 3. The abscissa of T ' is larger than or equal to the abscissa of T.

We have to show that: $\frac{\rho'}{N'} \geq \frac{\rho}{N}$. This follows immediately from (11).

Step 4. The absolute value of the slope of T'E is smaller than or equal to the slope of TE.

We have to prove that: $\frac{\frac{\sigma'-c'}{\sigma'}}{\frac{N'-\rho'}{N'}} \leq \frac{\frac{\sigma-c}{\sigma}}{\frac{N-\rho}{N}}$. This is equivalent with $\frac{N'}{\sigma'} \leq \frac{N}{\sigma}$ if σ' ≠ c'

and σ ≠ c. Again this inequality follows from (11). We already noticed that the case σ = c cannot occur.

Step 3 and 4 together show that T' is situated under or on the line segment TE.

This proves that in general L' is situated strictly under L (note that it is impossible that S = S' and T = T' because then $L_{r,s}$ and $R(L_{r',s'})$ would coincide).

Step 5. Exceptions

We now have a look at the exceptional cases. If ρ' = c' then S' = O, and hence by steps 3 and 4 the proposition holds (unless maybe if σ' = c'). Yet, if σ' = c' then $L_{r',s'}$ is the equality line, and again the inequality of the proposition holds strictly.

This proposition is the crucial step of the following theorem.

Theorem. If $L_{r,s}$ and $L_{r',s'}$ are intrinsically incomparable (where the sub-top precedes the top), then also $L_{r,s}$ and $R(L_{r',s'})$ are intrinsically incomparable.

This theorem implies that when studying Lorenz similarity it suffices to consider the case that the sub-top precedes the top, because if under these

circumstances the Lorenz similarity curves are incomparable, then considering a reflected curve will not make them comparable.

Proof. We have to show that if $L_{r,s}$ and $L_{r',s'}$ intersect, then $L_{r,s}$ and $R(L_{r',s'})$ also intersect. This statement is by contraposition equivalent with: if $L_{r,s}$ and $R(L_{r',s'})$ are intrinsically comparable then $L_{r,s}$ and $L_{r',s'}$ are intrinsically comparable. This is, in turn, equivalent with the expression: if $L_{r,s} \leq R(L_{r',s'})$ or $L_{r,s} \geq R(L_{r',s'})$ then $L_{r,s} \leq L_{r',s'}$ or $L_{r',s'} \leq L_{r,s}$.

If now $L_{r,s} \leq R(L_{r',s'})$ then $L_{r',s'} \geq R(L_{r,s})$ and the previous proposition implies that $L_{r',s'} \leq L_{r,s}$. If, $L_{r,s} \geq R(L_{r',s'})$, then $L_{r,s} \geq L_{r',s'}$ (immediately by the previous proposition). This proves the theorem.

Note

It is still possible that $L_{r,s} > L_{r',s'}$ (strictly) while $L_{r,s}$ and $R(L_{r',s'})$ are intrinsically incomparable. Indeed, consider the following example:

D = {r,s} with r = (1,1,1,0,0,0) and s = (0,1,1,1,1,1)

D'= {r', s'} with r' = (1,1,1,1,0,0) and s' = s.

Then the coordinates of S, T, S', T', S" and T" are:

$$S=\left(\frac{1}{6},\frac{1}{3}\right), T=\left(\frac{1}{2},\frac{3}{5}\right), S'=\left(\frac{1}{6},\frac{1}{4}\right), T'=\left(\frac{2}{3},\frac{2}{5}\right)$$

$$S''=\left(\frac{5}{6},\frac{1}{4}\right), T''=\left(\frac{1}{3},\frac{2}{5}\right)$$

Clearly (see Fig.3) $L_{r,s} > L_{r',s'}$ (strict) while $L_{r,s}$ and $R(L_{r',s'})$ intersect.

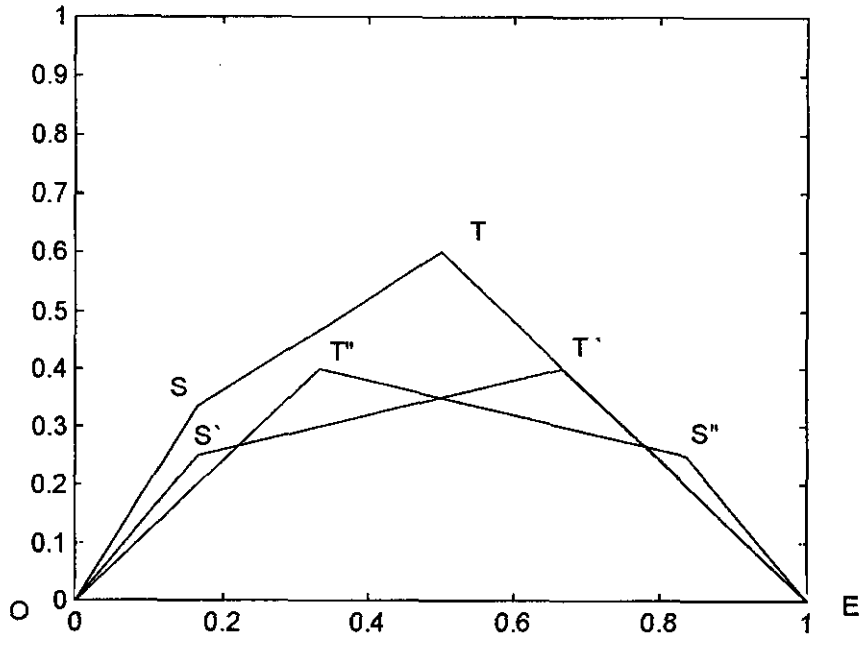

Fig.3 $L_{r,s} > L_{r',s'}$ (strictly) while $L_{r,s}$ and $R(L_{r',s'})$ intersect.

Considerations related to smallest and largest similarity

The equality line connecting the origin O = (0,0) with the point E = (1,0) is situated under all other Lorenz similarity curves. This line corresponds to the equivalence class with the largest similarity.

If the length, say N, of the keyword array is fixed, then the ordinate of the top is at most one, a case occurring when c = 0; its abscissa may be any of the values {1/N, 2/N, ..., (N-1)/N}. These abscissas lead to N-1 different (but pairwise equivalent) and intersecting curves. Consequently there is no minimum intrinsic

similarity curve, even when N is fixed (unless N = 2), and certainly not for variable N.

Theorem

Adding one keyword to the two keyword arrays of a duo (not equivalent to the equality line) such that the corresponding keyword is present, strictly increases similarity.

Proof. Without loss of generality we may consider the array r first and then s. Adding one 1 to the two N-arrays yields the following new difference array (of length N+1): $(|r| - c)$ times the value $(|r|+1)^{-1}$, $c+1$ times the value $((|r|+1)^{-1} - (|s|+1)^{-1})$, and $(|s| - c)$ times the value $-(|s|+1)^{-1}$.

If we assume that $|s| \geq |r|$ then $\sigma = |s|$ and $\rho = |r|$, leading to the situation schematically represented in Fig. 4. Using the notation S, T and S', T' as in Fig.2 we have to show that this figure correctly (although schematically) represents the transformation. This is: we will show that the curve O - S'-T'-E is situated under the curve O–S–T-E.

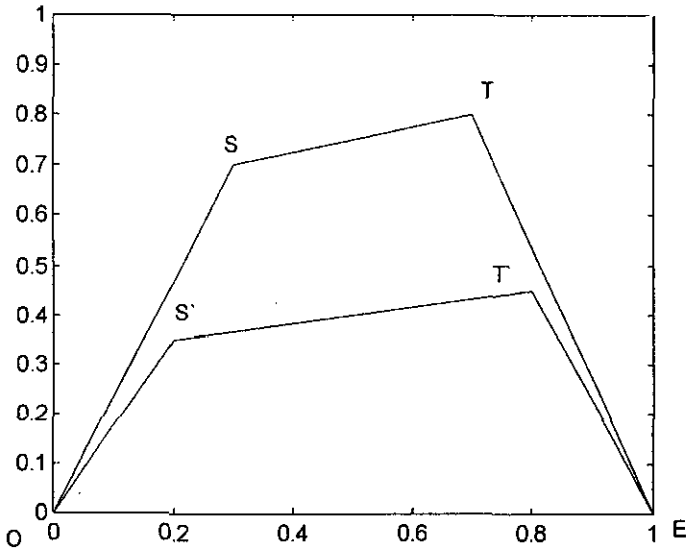

Fig.4 . Schematic representation (shifts are exaggerated) of the result of adding ones to two document representations

The coordinates of point S are $\left(\frac{\rho-c}{N},\frac{\rho-c}{\rho}\right)$, while point T has coordinates $\left(\frac{\rho}{N},\frac{\sigma-c}{\sigma}\right)$. After the transformation (i.e. adding ones to each array) we obtain the following points: S'= $\left(\frac{\rho-c}{N+1},\frac{\rho-c}{\rho+1}\right)$ and T' = $\left(\frac{\rho+1}{N+1},\frac{\sigma-c}{\sigma+1}\right)$. Clearly the abscissa of S' is smaller than that of S (unless ρ = c). As, moreover, the slope of the line segment OS', namely $\frac{N+1}{\rho+1}$, is strictly smaller than the slope of OS, $\frac{N}{\rho}$, (unless ρ = N) this proves that generally S' is situated below the line OS. Further, the abscissa of T' is clearly larger than that of T, unless ρ = N. The absolute value of the slope of TE is $\frac{N}{\sigma}$ (use the relation N = ρ + σ –c) , while the absolute value of the slope of TE' is $\frac{N+1}{\sigma+1}$. This proves that T' is situated below

the line segment TE. Finally, we have a look at the exceptions. First, if $\rho = c$ then $S = S' = O$, while T' is still situated under TE. Finally, if $\rho = N$, then also $\sigma = c = N$ and we are dealing with the equality line, a case that has been excluded.

Lorenz similarity functions

Definition

A Lorenz similarity function, f, is a real-valued function mapping a duo D to its Lorenz similarity value f(D). This function must, moreover, respect the Lorenz similarity partial order. This means that, if

$$D < D' \quad \text{then } f(D) < f(D')$$
$$\text{and if D and D" are equivalent then } f(D) = f(D")$$

In the case the function f only satisfies the requirement

$$D < D' \quad \text{then } f(D) \leq f(D')$$
$$\text{and if D and D" are equivalent then } f(D) = f(D")$$

then we say that f is a weak Lorenz similarity function.

These requirements imply that $f(\{r,s\}) = f(\{s,r\})$. Hence, expressing Lorenz similarity functions directly as functions of the difference array d, the requirements express that similarity functions must be symmetric, in the sense that $f(d) = f(-d)$. The approach taken here is closely related to the approach we took when studying so-called symmetric relative concentration (Egghe & Rousseau, 2001; Rousseau, 2001). Indeed, replication invariant Lorenz similarity curves are the same as the Lorenz curve variant used to study symmetric relative concentration. Yet, when concentration increases, similarity decreases and vice

versa. The point is that we can derive Lorenz similarity functions from the corresponding symmetric relative concentration functions. Recall that the ultimate proof that these functions respect the Lorenz similarity partial order is based on Egghe's general theory of concentration measures and their construction (2002).

In order to study the Jaccard measure (and the Gini similarity measure) we first prove the following lemma.

Lemma

The expression $-\frac{1}{N}\sum_{i=1}^{N} id_i$ is equal to the area under the Lorenz similarity curve.

Proof. The area under the Lorenz similarity curve is, using (2), equal to:

$$\frac{1}{2N}\left(c_1+\sum_{j=1}^{N-2}\left(c_j+c_{j+1}\right)+c_{N-1}\right)$$

$$=\frac{1}{2N}\left(d_1+(2d_1+d_2)+(2d_1+2d_2+d_3)+\right.$$

$$\ldots+(2d_1+\ldots+2d_i+d_{i+1})+$$

$$\left.\ldots+(2d_1+\ldots+2d_{N-2}+d_{N-1})+(d_1+d_2+\ldots+d_{N-1})\right)$$

$$=\frac{1}{N}\left((N-1)d_1+(N-2)d_2+\ldots+d_{N-1}\right)$$

$$=\frac{1}{N}\sum_{j=1}^{N-1}(N-j)d_j-\frac{1}{N}N\sum_{j=1}^{N}d_j \qquad \text{by (4)}$$

$$=-\frac{1}{N}\sum_{i=1}^{N}id_i$$

The Gini similarity measure

One of the best known concentration measures is the Gini index (Myles, 1995). It is easy to derive a Gini similarity measure, denoted as G_S, from the Gini concentration index:

$$G_S(D) = G_S\{r,s\} = 1 + \frac{2}{N} \sum_{i=1}^{N} id_i \qquad (12)$$

where the d_i are ranked in decreasing order. This is, by the lemma, nothing but one minus twice the area under the Lorenz similarity curve. This normalizes the Gini similarity measure in such a way that all minimal Lorenz similarity curves correspond to a Gini-value of zero, and the equality line has a similarity value of one. As shown in the next proposition, the Gini similarity measure coincides with the Jaccard index. Recall that in information retrieval the Jaccard index is equal to the number of keywords that the two items have in common, divided by the number of keywords used by at least one of the two items. In set-theoretic notation as used in the introduction of this article, it is the number of elements present in the intersection of sets A and B, divided by the number of elements in their union. Hence, with the notations introduced above, the Jaccard index J is equal to c/N. In citation analysis the relative co-citation (Egghe & Rousseau, 1990, p. 240) is nothing but the Jaccard index applied to presence-absence citation data.

Proposition

The Gini similarity measure for absence-presence data is equal to the Jaccard index, J.

Proof. We may assume, without loss of generality, that the array with the smaller number of ones is considered first. This gives:

$$-\frac{1}{N}\sum_{i=1}^{N} id_i = -\frac{1}{N}\left(\sum_{i=1}^{\rho-c}\frac{i}{\rho} + \sum_{i=\rho-c+1}^{\rho} i\left(\frac{1}{\rho}-\frac{1}{\sigma}\right) - \sum_{i=\rho+1}^{N}\frac{i}{\sigma}\right)$$

$$= -\frac{1}{N}\left(\frac{1}{\rho}\sum_{i=1}^{\rho} i - \frac{1}{\sigma}\sum_{i=\rho-c+1}^{N} i\right)$$

$$= -\frac{1}{N}\left(\frac{1}{\rho}.\frac{\rho.(\rho+1)}{2} - \frac{1}{\sigma}.\frac{N(N+1)}{2} + \frac{1}{\sigma}.\frac{(\rho-c).(\rho-c+1)}{2}\right)$$

$$= -\frac{1}{N}\left(\frac{1}{\rho}.\frac{\rho.(\rho+1)}{2} - \frac{1}{\sigma}.\frac{(\rho+\sigma-c)(\rho+\sigma-c+1)}{2} + \frac{1}{\sigma}.\frac{(\rho-c).(\rho-c+1)}{2}\right)$$

$$= -\frac{1}{N}\left(\frac{\rho+1}{2} - \frac{1}{\sigma}.\frac{\sigma(2\rho+\sigma-2c+1)}{2}\right)$$

$$= \frac{1}{N}\left(\frac{(\rho+\sigma-c)-c}{2}\right)$$

Consequently:

$$G_S = 1-2\left(-\frac{1}{N}\sum_{i=1}^{N} id_i\right) = 1-2\left(\frac{1}{N}\left(\frac{N-c}{2}\right)\right) = \frac{c}{N} = J \qquad (13)$$

Equality (13) shows that J is a Lorenz similarity measure as defined above.

In order to show how other classical retrieval and overlap measures fit into the Lorenz framework we study the overlap measures O_1 and O_2 (Egghe & Michel, 2002).

Theorem

The overlap measures $O_2 = \frac{c}{\sigma}$ and $O_1 = \frac{c}{\rho}$ are weak Lorenz similarity measures.

The proof of this theorem is based on the following lemma.

Lemma
If sub-top precedes top then the following two statements are true:

a) If the tops of two Lorenz similarity curves coincide then their sub-tops also coincide.

b) If the sub-tops of two Lorenz similarity measures coincide then their tops also do, except in the case that the sub-tops are in the origin O.

This lemma implies that (in most cases) if either the tops or the sub-tops of two Lorenz similarity curves coincide then the curves themselves coincide, a rather surprising result.

We first prove the lemma.

Clearly:

$$T=\left(\frac{\rho}{N},\frac{\sigma-c}{\sigma}\right)=\left(\frac{\rho-c}{N},\frac{\rho-c}{\rho}\right)+\left(\frac{c}{N},\frac{c}{\rho}-\frac{c}{\sigma}\right)=S+\left(\frac{c}{N},\frac{c}{\rho}-\frac{c}{\sigma}\right) \qquad (14)$$

We first show that if the tops coincide, this is $\left(\frac{\rho}{N},\frac{\sigma-c}{\sigma}\right)=\left(\frac{\rho'}{N'},\frac{\sigma'-c'}{\sigma'}\right)$ (*), then also the sub-tops coincide. For this to hold we have to show that $\left(\frac{c}{N},\frac{c}{\rho}-\frac{c}{\sigma}\right)=\left(\frac{c'}{N'},\frac{c'}{\rho'}-\frac{c'}{\sigma'}\right)$. From equality (*) we see that $\frac{\rho}{N}=\frac{\rho'}{N'}$, and $\frac{c}{\sigma}=\frac{c'}{\sigma'}$, hence $\frac{\sigma-c}{\rho}=\frac{\sigma'-c'}{\rho'}$. Dividing numerator and denominator by c (left-hand side, if

$c \neq 0$) and by c' (right-hand side, if $c' \neq 0$) yields $\frac{\frac{\sigma}{c}-1}{\frac{\rho}{c}} = \frac{\frac{\sigma'}{c'}-1}{\frac{\rho'}{c'}}$. From this equality and (*) we derive that $\frac{c}{\rho} = \frac{c'}{\rho'}$ (if $\sigma \neq c$, $\sigma' \neq c'$). This proves that the ordinates of S and S' coincide. Finally, $\frac{N}{c} = \frac{\rho+\sigma-c}{c} = \frac{\rho'+\sigma'-c'}{c'} = \frac{N'}{c'}$, showing that also the abscissas coincide. We will now consider the exceptional cases. If c = 0, then S = T and the ordinate of T = 1. As T = T' this implies that c'= 0, and hence also S' = T', and the two curves coincide. Similarly, if c' = 0, then from T'= T, we see that c = 0 too, and thus S = T. If $\sigma = c$ and $\sigma' = c'$, then, since $c \leq \rho \leq \sigma$, and $c' \leq \rho' \leq \sigma'$ we conclude that $\sigma = \rho = c$ and $\sigma' = \rho' = c'$. This means that we have the equality curve in both cases.

If the sub-tops coincide, this is $\left(\frac{\rho-c}{N}, \frac{\rho-c}{\rho}\right) = \left(\frac{\rho'-c'}{N'}, \frac{\rho'-c'}{\rho'}\right)$ (**), then, in order to prove that also the tops coincide, we again have to show that $\left(\frac{c}{N}, \frac{c}{\rho}-\frac{c}{\sigma}\right) = \left(\frac{c'}{N'}, \frac{c'}{\rho'}-\frac{c'}{\sigma'}\right)$. Now from the equality $\frac{\rho-c}{N} = \frac{\rho'-c'}{N'}$ (by **) we derive: $\frac{\rho+\sigma-c}{\rho-c} = \frac{\rho'+\sigma'-c'}{\rho'-c'}$ (if $\rho \neq c$, and $\rho' \neq c'$); hence also $\frac{\sigma}{\rho-c} = \frac{\sigma'}{\rho'-c'}$. Dividing numerator and denominator by c (left-hand side, if $c \neq 0$) and by c' (right-hand side, if $c' \neq 0$) yields: $\frac{\frac{\sigma}{c}}{\frac{\rho}{c}-1} = \frac{\frac{\sigma'}{c'}}{\frac{\rho'}{c'}-1}$, and hence, again by (**): $\frac{\sigma}{c} = \frac{\sigma'}{c'}$. This result, together with (**) proves that the ordinates of the tops coincide. Finally:

$\frac{N}{c} = \frac{\rho+\sigma-c}{c} = \frac{\rho'+\sigma'-c'}{c'} = \frac{N'}{c'}$, proving that also the abscissas of the tops coincide.

We will now consider the exceptional cases. If c = 0, then S = T and the ordinate of T = 1. As S = S' this implies that c'= 0, and hence also S' = T' and the two curves coincide. Similarly, if c' = 0, then S'= S, also c = 0 and thus S = T. If ρ = c, then S = O, hence also S'= O. Under these circumstances it is still possible that T and T' are different. An example of this situation is D ={ r = (1,1,1,0), s = (1,1,1,1)} and D'= {r'= (1,1,0,0), s' = (1,1,1,1)}. Here S = S' = O, but T = (3/4,1/4), while T'= (2/4,2/4).

This proves the lemma.

Proof of the theorem

Assume that D'= {r',s'} > D = {r,s}. The Lorenz curve of D (respectively D') will be denoted by L (respectively L'). The relation D' > D means that either L is situated strictly above L' or that L is situated strictly above R(L'), the reflection of L' with respect to the line x = 1 / 2 .

If the ordinate of T (the top of the D-curve) is strictly larger than the ordinate of the top of D' (denoted as T') then clearly, $O_2' < O_2$. If, however, the ordinates of the two tops coincide, then the curves should be equivalent. Yet, it is possible that one Lorenz curve is situated strictly under the other while the ordinates of the tops coincide. This may happen when one curve has a horizontal segment.

An example of this situation is D = {(1,1,1,1,0,0),(0,0,1,1,1,1)} and D'= {(1,1,1,0,0,0), (1,1,1,1,1,1)}. D' is strictly below D, but the ordinates of their tops coincide. Recall that a horizontal segment only occurs if $\rho = \sigma$, and $c \neq 0$. This proves that O_2 is only a weak Lorenz similarity measure.

We next consider the overlap measure O_1. It suffices to consider the case that L is situated above L' (by the theorem proved above the case that L is situated above R(L') must not be considered). Then the absolute value of the slope of TE is N/σ, while the absolute value of the slope of T'E is N'/σ'. Then $\frac{N'}{\sigma'} \leq \frac{N}{\sigma}$ (because L is situated above L'), and consequently $\frac{\rho'-c'}{N'} \leq \frac{\rho-c}{N}$, or: the abscissa of S' ≤ abscissa of S. Again, because L is situated above L', this implies that the ordinate(S') ≤ ordinate(S) . If ordinate(S') = ordinate(S) and abscissa of S' ≤ abscissa (S) then automatically S'= S, and hence, by the lemma the two Lorenz curves coincide, unless S' = S = O. If ordinate(S') is not equal to ordinate(S), then ordinate(S') < ordinate(S), and hence $O_1' > O_1$. So, in general we can only say that D' > D implies $O_1' \geq O_1$. This proves that the overlap measure O_1 is a weak Lorenz similarity measure.

The relation between retrieval and overlap measures, and Lorenz similarity

We first show that the cases where strictly different Lorenz similarity measures lead to the same overlap value do not co-occur for O_1 and O_2.

Proposition (using the same notation as above)

Assume that D' > (strictly) D, then either $O_1' > O_1$ or $O_2' > O_2$.

Proof. We know already that under this assumption $O_1' \geq O_1$ and $O_2' \geq O_2$. If the conclusion of this proposition were not correct then $O_1' = O_1$ and $O_2' = O_2$. Under this assumption $\frac{c}{\rho} = \frac{c'}{\rho'} \& \frac{c}{\sigma} = \frac{c'}{\sigma'}$ implying that $\frac{N}{c} = \frac{N'}{c'}$. Now three cases are possible:

$$(I)\ \frac{\rho - c}{N} < \frac{\rho' - c'}{N'} \Leftrightarrow \frac{\rho}{N} < \frac{\rho'}{N'}$$

$$(II)\ \frac{\rho - c}{N} > \frac{\rho' - c'}{N'} \Leftrightarrow \frac{\rho}{N} > \frac{\rho'}{N'}$$

$$(III)\ \frac{\rho - c}{N} = \frac{\rho' - c'}{N'} \Leftrightarrow \frac{\rho}{N} = \frac{\rho'}{N'}$$

In the third case S=S' and T=T', hence the curves coincide, which is excluded by the strict inequality D' > D. In the two other cases the Lorenz curves intersect, and hence we have a contradiction. This shows that it is impossible that $O_1' = O_1$ and $O_2' = O_2$. This proves the proposition.

Because we consider this to be an important result, we provide a second proof.

The difference between the abscissas of top and sub-top is always c/N, which is also equal to the Jaccard index. If now D' > (strictly) D, then J' > J, and hence the difference between the abscissas of D' is larger than that between the abscissas of D. If $O_2' = O_2$ then S' must be situated to the left of S (even if T and T'

coincide). Hence the ordinate of S' is strictly smaller than that of S, or $O_1' > O_1$. If $O_1' = O_1$ and from the fact that D'> D, if follows that S must be situated to the left of S' (or that they coincide). Then T must be situated strictly above T', or $O_2' > O_2$.

Corollary

Any average of the overlap measures O_1 and O_2 is an acceptable Lorenz (non-weak!) similarity measure.

Taking the harmonic mean leads to the well-known Dice coefficient. Indeed, the harmonic mean of O_1 and O_2 is:

$$\frac{2}{\frac{1}{O_1}+\frac{1}{O_2}} = \frac{2c}{\rho+\sigma} \tag{15}$$

Note that in an information retrieval context, ρ + σ denotes the number of keywords in document representation r plus the number of keywords in document representation s; c is the number of keywords they have in common.

The geometric mean of O_1 and O_2 is Salton's cosine measure:

$$\sqrt{\frac{c}{\rho}\cdot\frac{c}{\sigma}} = \frac{c}{\sqrt{\rho.\sigma}} \tag{16}$$

Hence this is also an acceptable measure in our framework. We further note that all basic measures of information retrieval (precision, recall, fallout and miss) (Egghe, 2004) can be expressed using the overlap measures O_1 and O_2, but they are all weak similarity measures in our sense.

Another example of an acceptable similarity measure is

$$V_S^2(D)=V_S^2\{r,s\} = \frac{1}{N\sum_{i=1}^{N} d_i^2} \tag{17}$$

This is the adapted Simpson or Herfindahl index of the relative difference vector. It is related to the squared coefficient of variation, hence the notation V_S^2. The factor $\sum_{i=1}^{N} d_i^2$ is equal to $\frac{(\rho-c)+(\sigma-c)}{\rho.\sigma}$, hence in an information retrieval setting it can be interpreted as the number of unique keywords (keywords that are either unique to r or unique to s) divided by the product of the number of keywords in r with the number of keywords in s. In this form the adapted Simpson index has the drawback that if r = s the similarity value becomes $+\infty$. The similarity of minimal Lorenz curves tend to zero as N increases.

We finally note that any increasing function of a Lorenz similarity function is again an acceptable Lorenz similarity function.

Conclusion

We have shown that classical measures used in information retrieval, studies of indexer consistency and overlap studies can be characterized by Lorenz similarity curves. This provides a visual, geometric picture of similarity, different from the geometric approach based on iso-similarity curves, as studied by Jones and Furnas (1987). We have shown that the Jaccard index, the Dice coefficient and Salton's cosine measure respect the partial order determined by these Lorenz similarity curves, hereby revealing new good properties of these important measures. Our approach explains how, at least formally, the information sciences can be linked to the big economic and social theories where the Lorenz curve and derived measures are basic tools (Atkinson, 1970; Dalton, 1920; Myles, 1995). From a conceptual point of view the importance of such a relation cannot be underestimated.

REFERENCES

Atkinson, A.B. (1970). On the measurement of inequality. *Journal of Economic Theory, 2,* 244-263.

Dalton, H. (1920). The measurement of the inequality of incomes. *The Economic Journal, 30,* 348-361.

Egghe, L. (2002). Construction of concentration measures for general Lorenz curves using Riemann-Stieltjes integrals. *Mathematical and Computer Modelling. 35,* 1149-1163.

Egghe, L. (2004). A universal method of information retrieval evaluation: the "missing" link M and the universal IR surface. *Information Processing and Management, 40*, 21-30.

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

Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics*. Amsterdam: Elsevier.

Egghe, L., & Rousseau, R. (2001). Symmetric and asymmetric theory of relative concentration and applications. *Scientometrics, 52*, 261-290.

Huot, C., Quoniam, L., & Dou, H. (1992). A new method for analyzing downloaded data for strategic decision. *Scientometrics, 25*, 279-294.

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

Lorenz, M.O. (1905). Methods of measuring concentration and wealth. *Journal of the American Statistical Association, 9*, 209-219.

Magurran, A.E. (1991). *Ecological diversity and its measurement*. London: Chapman & Hall.

Myles, Gareth D. (1995). *Public Economics*. Cambridge: Cambridge University Press.

Rousseau, R. (2001). Concentration and evenness measures as macro-level scientometric indicators. In: *Keyan pingjia yu daxue pingjia (Research evaluation and university evaluation)*, (Wang, Z., Jiang, J., eds.). Beijing: Red Flag Publishing House, 72-89. (In Chinese, an English translation is available from the author).

Salton G., & McGill, M.J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.