Symmetric and Asymmetric Theory of Relative Concentration and Applications

EGGHE, Leo & ROUSSEAU, Ronald (2001) Symmetric and Asymmetric Theory of Relative Concentration and Applications. In: Scientometrics, 52(2). p. 261-290.

DOI: 10.1023/A:1017967807504
Handle: http://hdl.handle.net/1942/780

SYMMETRIC AND ASYMMETRIC THEORY OF RELATIVE CONCENTRATION AND APPLICATIONS

by

L. Egghe — LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
UIA, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
e-mail : leo.egghe@luc.ac.be

R. Rousseau — KHBO, Department of Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende, Belgium[1]
and
UIA, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
e-mail : ronald.rousseau@kh.khbo.be

Abstract

Relative concentration theory studies the degree of inequality between two vectors $(a_1,\ldots,a_N)$ and $(\alpha_1,\ldots,\alpha_N)$. It extends concentration theory in the sense that, in the latter theory, one of the above vectors is $(\frac{1}{N},\ldots,\frac{1}{N})$ (*N* coordinates).

When studying relative concentration one can consider the vectors $(a_1,\ldots,a_N)$ and $(\alpha_1,\ldots,\alpha_N)$ as interchangeable (equivalent) or not. In the former case this means that the relative concentration of $(a_1,\ldots,a_N)$ versus $(\alpha_1,\ldots,\alpha_N)$ is the same as the relative concentration of $(\alpha_1,\ldots,\alpha_N)$ versus $(a_1,\ldots,a_N)$. We deal here with a symmetric theory of relative concentration. In the other case one wants to consider $(a_1,\ldots,a_N)$ as having a different role

---

[1]Permanent address

Key words and phrases : relative concentration, symmetric, asymmetric, concentration measure, Lorenz, Gini, Pratt

as $(\alpha_1,\ldots,\alpha_N)$ and hence the results can be different when interchanging the vectors. This leads to an asymmetric theory of relative concentration.

In this paper we elaborate both models. As they extend concentration theory, both models use the Lorenz order and Lorenz curves.

For each theory we present good measures of relative concentration and give applications of each model.

I. Introduction.

Classical concentration theory deals with a vector $(x_1,\ldots,x_N)$ of numbers x_i. In most cases $x_i \geq 0$ for $i=1,\ldots,N$. Concentration theory is a theory of how to measure inequality between the numbers $x_1,\ldots,x_N$. It originates from econometrics where it was used to measure the inequality of wealth in a population (a social group, a country, ...). One of the earliest concentration measures is the so-called Gini index, Gini (1909), later re-invented by Pratt (1977) in informetrics - see Carpenter (1979). Concentration theory has since then been studied in the context of informetric problems - see Egghe and Rousseau (1990a, 1990b, 1991) and Rousseau (1992).

By its very definition, studying inequality (or concentration) of a vector $(x_1,\ldots,x_N)$ is comparing its relative scores $(a_1,\ldots,a_N)$, where

$$a_i = \frac{x_i}{\sum_{j=1}^{N} x_j}, \qquad (1)$$

with the vector of equality $(\frac{1}{N},\ldots,\frac{1}{N})$ (N coordinates). In the above publications this is done by constructing the so-called Lorenz curve of the vector $(x_1,\ldots,x_N)$ or $(a_1,\ldots,a_N)$, namely by connecting in the plane the points (0,0) and

$$\left(\frac{i}{N}, \sum_{j=1}^{i} a_j\right) \qquad (2)$$

for all $i=1,\ldots,N$ (note that for $i=N$, (2) becomes (1,1)). Here it is understood that the a_is are arranged monotonically, either increasing or decreasing (yielding two equivalent models). In this paper we will restrict ourselves to decreasing orders. In this case, a Lorenz curve looks like the one in Fig. 1.

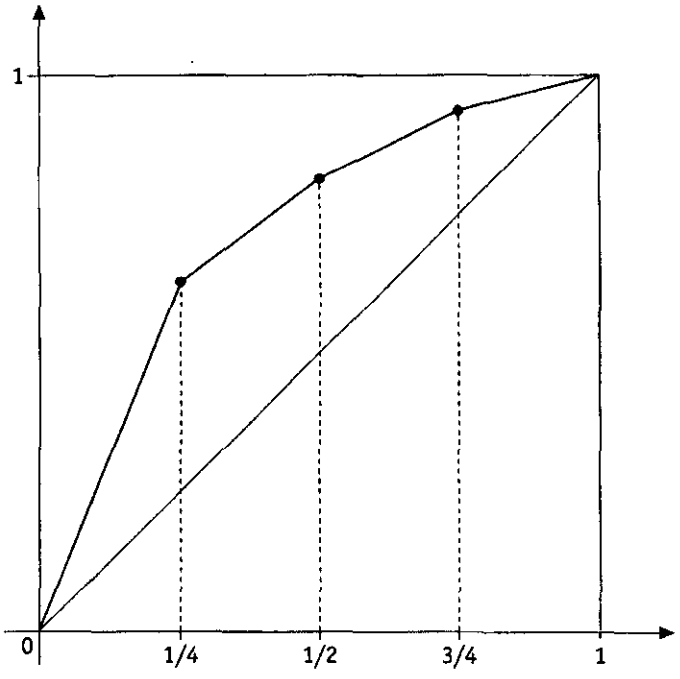

Fig. 1 Example of a Lorenz curve (N=4)

It can be shown (but it is also intuitively appealing) that, if we have two vectors $X=(x_1,\ldots,x_N)$ and $Y=(y_1,\ldots,y_N)$ such that the Lorenz curve of X is below the one of Y, that the vector Y represents a more concentrated situation. If this is the case we write $X---<Y$. It is then clear that a good concentration measure must respect this order. Therefore we define

$$f : \mathbb{R}^N \to \mathbb{R}$$
$$X=(x_1,\ldots,x_N) \to f(X)=f(x_1,\ldots,x_N) \tag{3}$$

to be a good concentration measure if $X---<Y$ and $X \neq Y$ imply $f(X)<f(Y)$. Many examples of good concentration measures can be given. We restrict ourselves to Pratt's measure (or

the Gini index), the coefficient of variation and Theil's measure. To start with the coefficient of variation V is just

$$V = \frac{\sigma}{\mu}, \tag{4}$$

where σ is the standard deviation and μ is the average of $X=(x_1,\ldots,x_N)$. When we rank X in decreasing order, we have that Pratt's measure is given by

$$C = \frac{2\left(\frac{N+1}{2} - q\right)}{N-1} \tag{5}$$

where

$$q = \sum_{i=1}^{N} i a_i \ .$$

Gini's index is then

$$G = \frac{N-1}{N} C \ . \tag{6}$$

Note that G is twice the area between the Lorenz curve of X and the diagonal of the unit square connecting (0,0) and (1,1), i.e. the Lorenz curve of the vector of equality $(\frac{1}{N},\ldots,\frac{1}{N})$. *Equation (6) in fact shows that Pratt re-invented G, as explained in Carpenter* (1979).

Finally, Theil's measure is given by (Theil (1967))

$$Th = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{\mu} \log\left(\frac{x_i}{\mu}\right) \ . \tag{7}$$

The problem that *we will address in this paper is extending concentration theory (or theory* of inequality) to the case of comparing two vectors $X=(x_1,\ldots,x_N)$ and $Y=(y_1,\ldots,y_N)$, in the sense that we want to develop measures of inequality between these vectors. The intuitive idea is that - to mention only two extreme cases - if X=Y, the measure should be zero

while if, say X=(1,0) and Y=(0,1), the measure of inequality should be maximal. We will go into this further on. Let us first of all indicate that such a theory of relative concentration - besides its mathematical challenge - is needed in practise, as will be explained now.

Take the case of information retrieval (IR) where documents and queries are represented by vectors of the type $X=(x_1,\ldots,x_N)$. Here x_i is the weight of key word i in this document or query - see e.g. Salton and Mc Gill (1987). If we have a measure of relative concentration we are able to give values to the relative similarity (i.e. the opposite of inequality, concentration or dissimilarity) of two documents or of two queries or, most importantly, between a document and a query. Documents with high similarity values can be retrieved based on a threshold that can be put on these values. This method has been applied in Egghe (1990). Certainly, when comparing two documents d_1 and d_2 or two queries q_1 and q_2, their roles are symmetrical in the sense that the relative inequality between d_1 and d_2 is the same as the relative inequality between d_2 and d_1. The same goes for q_1 and q_2. Hence here we need symmetric relative concentration theory. In the case of comparison of a query q with a document d one could be inclined to say that now there is asymmetry since we are dealing with different objects. Then asymmetric relative concentration theory is needed. But in the light of a dual vision between queries and documents (see Egghe and Rousseau (1997)) one could also argue for applying symmetrical measures.

Another application is given in Viles and French (1999). The authors want to find a measure of content locality in distributed document collections or in digital libraries. Digital libraries are here considered as a set of autonomous, distinct document collections (so-called sites) that cooperate to support IR. The relative sizes of these sites constitute a first vector X. Next, one is interested in how a certain topic is distributed over these sites. This gives a second vector Y and topic locality is nothing else than the relative concentration between X and Y. Some remarks are in order. First of all, since one is interested in the topic locality, it is our feeling that an asymmetric measure of relative concentration is needed. Indeed, we do not compare different topics but all topics are compared with X, the distribution of the sizes of the sites. We will go into this further on but it is our feeling that Viles and French attempt a symmetric approach. But their measure (see later) is not fully explained. Indeed,

only fragmentary properties are described and proofs are hardly given and only in the case that all the sites have the same size. In view of the above they only deal with concentration, but even this (well-known) theory is not used. We will discuss this further on, once we have established our relative concentration models.

In the next section we will introduce a theory of symmetric relative concentration. We will include measures that can be used in this connection. The measure used in Viles and French (1999) appears here (the measure of relative variation) but also the relative Pratt measure as was introduced already in Egghe (1988), improving Pratt's own relative concentration measure.

The third section is devoted to the study of asymmetric relative concentration. Here we use the notion of weighted concentration (see Rousseau (1992)) which is interpreted (explained) as overlap concentration (overlap in the sense of collection overlap).

The last section is devoted to applications of these measures. It is shown that the classical cosine formula is not a good measure of symmetric relative concentration. We also include a discussion of the measures that are needed in Viles and French (1999) to describe content locality, hereby giving more fundamental support for the experiments that are done in that paper. We argue that Viles and French better use asymmetric relative concentration measures instead of the symmetric one that is used now.

II. Symmetric relative concentration theory.

Let us consider two vectors $X=(x_1,\ldots,x_N)$ and $Y=(y_1,\ldots,y_N)$ with $N\in\mathbb{N}$ fixed. Here the x_is and y_is are real numbers but in practical applications these numbers will be positive. Denote by

$$a_i = \frac{x_i}{\sum_{j=1}^{N} x_j} \tag{8}$$

$$\alpha_i = \frac{y_i}{\sum_{j=1}^{N} y_j} \tag{9}$$

the relative scores. Vectors of relative scores are denoted as A_X and A_Y respectively. If we want to compare X and Y in a symmetric way we will, in analogy with concentration theory, work with the vectors A_X and A_Y and study the difference $A_X - A_Y = (a_i - \alpha_i)_{i=1,\dots,N}$.

Let f_r be any function

$$f_r : \mathbb{R}^N x \mathbb{R}^N \to \mathbb{R}$$
$$(X,Y) \to f_r(X,Y) \tag{10}$$

II.1. Symmetric relative concentration axiom and properties.

Definition II.1.1 : We say that f_r is a good measure of symmetric relative concentration if f is symmetric and if for any two situations (X,Y) and (X′,Y′) such that

$$A_X - A_Y \text{ ---}< A_{X'} - A_{Y'} \tag{11}$$

and such that $A_X - A_Y \neq A_{X'} - A_{Y'}$, we have that

$$f_r(X,Y) < f_r(X',Y') \tag{12}$$

Here ---< is the Lorenz order, meaning that the Lorenz curve of $A_X - A_Y$ is never above the Lorenz curve of $A_{X'} - A_{Y'}$. It is trivial to see that this is equivalent with (cf.(2))

$$\begin{cases} a_1 - \alpha_1 \le a_1' - \alpha_1' \\ (a_1+a_2) - (\alpha_1+\alpha_2) \le (a_1'+a_2') - (\alpha_1'+\alpha_2') \\ \vdots \\ \sum_{j=1}^{N-1} a_j - \sum_{j=1}^{N-1} \alpha_j \le \sum_{j=1}^{N-1} a_j' - \sum_{j=1}^{N-1} \alpha_j' \end{cases} \tag{13}$$

Here we have ordered A_X-A_Y and $A_{X'}$-$A_{Y'}$ decreasingly as assumed in this paper (we kept the notation for the indices, however, for reasons of simplicity).

Note that from definition II.1.1 it already follows that f_r attains its smallest value in case $X=Y$ (or, more generally, in case $A_X=A_Y$) and that f_r attains its largest value in cases where $A_X=(\delta_{ik})_{k=1,\ldots,N}$ and $A_Y=(\delta_{jk})_{k=1,\ldots,N}$ where $i\neq j$ (i.e. X and Y differ maximally). This is readily seen by drawing the Lorenz curves. Here δ denotes the Kronecker delta which takes the value 1 if the indices are equal and 0 if they are unequal.

Proposition II.1.2 : For any two situations (X,Y) and (X′,Y′) we have that

$$A_X - A_Y \ \text{---<}\ A_{X'} - A_{Y'} \tag{11}$$

iff

$$A_Y - A_X \ \text{---<}\ A_{Y'} - A_{X'} \tag{14}$$

Proof : By interchanging the roles of X and Y it suffices to prove only one of these implications. Suppose (11). This is equivalent with

$$\begin{cases} \alpha_1 - a_1 \geq \alpha'_1 - a'_1 \\ \cdot \\ \cdot \\ \cdot \\ \sum_{j=1}^{N-1} \alpha_j - \sum_{j=1}^{N-1} a_j \geq \sum_{j=1}^{N-1} \alpha'_j - \sum_{j=1}^{N-1} a'_j \end{cases} \tag{15}$$

Since

$$\sum_{j=1}^{N} a_j - \sum_{j=1}^{N} \alpha_j = \sum_{j=1}^{N} a'_j - \sum_{j=1}^{N} \alpha'_j = 0 \tag{16}$$

we have that (15) is equivalent with

$$\begin{cases} \sum_{j=2}^{N} a_j - \sum_{j=2}^{N} \alpha_j \geq \sum_{j=2}^{N} a'_j - \sum_{j=2}^{N} \alpha'_j \\ \cdot \\ \cdot \\ \cdot \\ a_N - \alpha_N \geq a'_N - \alpha'_N \end{cases} \tag{17}$$

$$\Leftrightarrow \quad \begin{cases} \alpha_N - a_N \leq \alpha'_N - a'_N \\ \cdot \\ \cdot \\ \cdot \\ \sum_{j=2}^{N} \alpha_j - \sum_{j=2}^{N} a_j \leq \sum_{j=2}^{N} \alpha'_j - \sum_{j=2}^{N} a'_j \end{cases} \tag{18}$$

Now since $(a_i-\alpha_i)_{i=1,\ldots,N}$ decreases, the same is true for $(\alpha_{N-i+1} - a_{N-i+1})_{i=1,\ldots,N}$. This together with (18) proves (14). □

Corollary II.1.3 : For any pair of vectors (X,Y), the Lorenz curve of A_Y-A_X coincides with the reflection of A_X-A_Y with respect to the vertical line $x=\frac{1}{2}$.

Proof : Abscissae have the form $\frac{i}{N}$, i=1,...,N. Hence the point that is symmetrical to $\frac{1}{2}$ w.r.t. $\frac{i}{N}$ is $\frac{N-i}{N}$. The value of the Lorenz curve of A_X-A_Y in $\frac{i}{N}$ is $\sum_{j=1}^{i} a_j - \sum_{j=1}^{i} \alpha_j$. According to the above proof, the value of the Lorenz curve of A_Y-A_X in $\frac{N-i}{N}$ is (see (18))

$$\sum_{j=i+1}^{N} \alpha_j - \sum_{j=i+1}^{N} a_j = \sum_{j=1}^{i} a_j - \sum_{j=1}^{i} \alpha_j,$$

again using (16). So the Lorenz curves of A_X-A_Y and of A_Y-A_X are symmetric w.r.t. $x=\frac{1}{2}$ in the points $\frac{i}{N}$, i=1,...,N. Hence connecting these points yields symmetric Lorenz curves.□

The above result can be illustrated as follows. Let X be such that $A_X=(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ and Y such that $A_Y=(\frac{1}{8}, \frac{1}{8}, \frac{3}{4})$. Then $A_X-A_Y=(\frac{3}{8}, \frac{1}{8}, -\frac{1}{2})$ and $A_Y-A_X=(\frac{1}{2}, -\frac{1}{8}, -\frac{3}{8})$ (decreasing order). The Lorenz curves are shown in Fig. 2

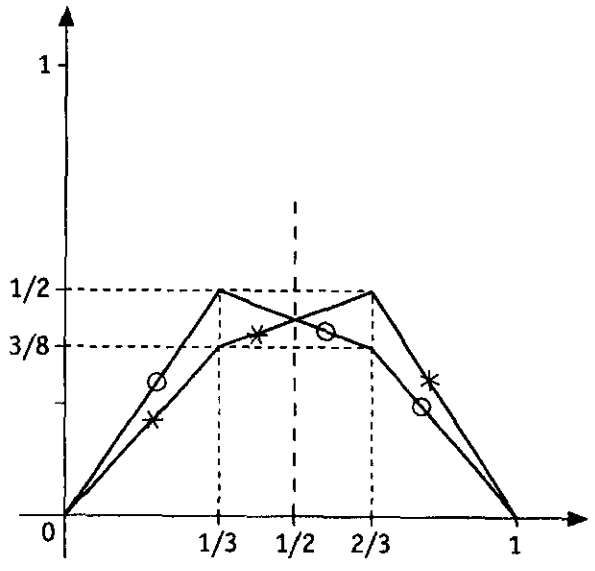

Fig. 2 Lorenz curves of A_X-A_Y (*) and A_Y-A_X(°).

Note that proposition II.1.2 is the reason why we want to have symmetric functions f_r as good measures of symmetric relative concentration.

The following results show that if we restrict our relative concentration theory to the case where one vector is constant, we obtain concentration theory.

Proposition II.1.4 :

1. Let $A_X=A_{X'}=(\frac{1}{N},...,\frac{1}{N})$ (N coordinates) and $A_Y=(\alpha_i)_{i=1,...,N}$, $A_{Y'}=(\alpha_i')_{i=1,...,N}$. We have

$$A_X - A_Y \ \text{---<}\ A_{X'} - A_{Y'}$$

iff Y ---< Y'.

2. Let $A_Y=A_{Y'}=(\frac{1}{N},\ldots,\frac{1}{N})$ (N coordinates) and $A_X=(a_i)_{i=1,\ldots,N}$, $A_{X'}=(a_i')_{i=1,\ldots,N}$. We have

$$A_X - A_Y \ \text{---<}\ A_{X'} - A_{Y'}$$

iff X ---< X'.

Proof : Upon switching the roles of X and Y and using proposition II.1.2 it suffices to prove only the second assertion. Now

$$A_X - A_Y \ \text{---<}\ A_{X'} - A_{Y'}$$

iff

$$(a_i - \frac{1}{N})_{i=1,\ldots,N} \ \text{---<}\ (a_i' - \frac{1}{N})_{i=1,\ldots,N}$$

iff (decreasing order)

$$\sum_{j=1}^{i}(a_j - \frac{1}{N}) \le \sum_{j=1}^{i}(a_j' - \frac{1}{N})$$

for all $i=1,\ldots,N-1$. This is equivalent with

$$\sum_{j=1}^{i} a_j \le \sum_{j=1}^{i} a_j'$$

for all $i=1,\ldots,N-1$. Now $(a_i)_{i=1,\ldots,N}$ and $(a_i')_{i=1,\ldots,N}$ decrease iff $(a_i-\frac{1}{N})_{i=1,\ldots,N}$ and $(a_i'-\frac{1}{N})_{i=1,\ldots,N}$ decrease. Hence A_X-A_Y ---< $A_{X'}$-$A_{Y'}$ is equivalent with X ---< X'. □

Corollary II.1.6 : if f_r is a good measure of symmetric relative concentration then $f_r|_D=f$ is a good measure of concentration. Here $f_r|_D$ denotes the restriction of f_r to the set D, where D is the set

$$\{(X,Y) \| Y=(\frac{1}{N},\ldots,\frac{1}{N}),\ X\in\mathbb{R}^N\} \tag{19a}$$

or the set

$$\{(X,Y) \| X=(\frac{1}{N},\ldots,\frac{1}{N}),\ Y\in\mathbb{R}^N\} \tag{19b}$$

Proof : This follows readily from the above proposition and the definitions of good measures of symmetric relative concentration and of concentration. □

As is the case in concentration theory, the above theory can only have value if there exist concrete measures f_r satisfying the above properties. This will be examined in the next subsection.

II.2 Measures of symmetric relative concentration.

In Pratt (1977) there was an attempt to define a measure of relative concentration. The formula is

$$\frac{1}{N-1}\sum_{i=1}^{N} i(a_i-\alpha_i), \tag{20}$$

where $A_X=(a_i)_{i=1,\ldots,N}$ and $A_Y=(\alpha_i)_{i=1,\ldots,N}$ as above. It was already noted in Egghe (1988) that this measure is not extreme in the extreme situations (i.e. of relatively equal or relatively opposite vectors. We also see immediately that the function in (20) is not even, a necessity for good symmetric relative concentration measures. The remedy for (20) is given in Egghe (1988) (we multiply by 2 for reasons which will become clear later on) :

$$C_r(X,Y) = \frac{2}{N-1}\max_{\varphi\in\pi_N}\sum_{i=1}^{N}\varphi(i)(a_i-\alpha_i), \tag{21}$$

where π_N denotes the set of all permutations of $\{1,\ldots,N\}$.

However, at the time of the writing of Egghe (1988) we did not have the requirements developed in the previous subsection : in Egghe (1988) we only made sure that C_r is extreme in the extreme cases. This is however implied by definition II.1.1, which we will check now.

Proposition II.2.1 : C_r is a good measure of symmetric relative concentration.

Proof : First we note that C_r is symmetric. Indeed, let $\nu \in \pi_N$ be such that

$$C_r(X,Y) = \frac{2}{N-1} \sum_{i=1}^{N} \nu(i)(a_i - \alpha_i).$$

Now

$$C_r(Y,X) \geq \frac{2}{N-1} \sum_{i=1}^{N} \xi(i)(\alpha_i - a_i)$$

for any $\xi \in \pi_N$. This is in particular the case for

$$\xi(i) = N - \nu(i) + 1$$

for every $i \in \{1,\ldots,N\}$. Note that $\xi \in \pi_N$. Now

$$\sum_{i=1}^{N} \xi(i)(\alpha_i - a_i) - \sum_{i=1}^{N} \nu(i)(a_i - \alpha_i)$$

$$= \sum_{i=1}^{N} [N - \nu(i) + 1 + \nu(i)] (\alpha_i - a_i)$$

$$= 0$$

Since $\sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} a_i = 0$. Hence

$$C_r(Y,X) \geq C_r(X,Y).$$

The other inequality follows by reversing the roles of X and Y. Hence C_r is symmetric. This result was already proved in Egghe (1988). All we have to do now is to verify if

$$A_X - A_Y \text{ ---} < A_{X'} - A_{Y'} \tag{22}$$

and if they are unequal, we have

$$C_r(X,Y) < C_r(X',Y') . \tag{23}$$

Suppose we have arranged A_X-A_Y and $A_{X'}$-$A_{Y'}$ in decreasing order, as required for the Lorenz curves. Then (21) reduces to

$$C_r(X,Y) = \frac{2}{N-1} \sum_{i=1}^{N} (N-i+1)(a_i-\alpha_i)$$

$$C_r(X',Y') = \frac{2}{N-1} \sum_{i=1}^{N} (N-i+1)(a_i'-\alpha_i')$$

Now (22) is equivalent with

$$a_1-\alpha_1 \le a_1'-\alpha_1'$$

$$(a_1+a_2) - (\alpha_1+\alpha_2) \le (a_1'+a_2') - (\alpha_1'+\alpha_2')$$

.

.

.

$$\sum_{i=1}^{N-1} a_i - \sum_{i=1}^{N-1} \alpha_i \le \sum_{i=1}^{N-1} a_i' - \sum_{i=1}^{N-1} \alpha_i'$$

$$\sum_{i=1}^{N} a_i - \sum_{i=1}^{N} \alpha_i = \sum_{i=1}^{N} a_i' - \sum_{i=1}^{N} \alpha_i' = 0$$

If we add all these lines we get

$$\sum_{i=1}^{N} (N-i+1)(a_i-\alpha_i) < \sum_{i=1}^{N} (N-i+1)(a_i'-\alpha_i')$$

($<$ since at least one of the inequalities is strict, since $A_X-A_Y \neq A_{X'}-A_{Y'}$) This proves (23). □

From this result we know that $C_r|_D$ (D as in corollary II.1.6) is a good measure of concentration. This can indeed be seen directly from the next proposition.

Proposition II.2.2 : $C_r|_D = C$ = Pratt's measure of concentration.

Proof : By the fact that C_r is a good measure of symmetric relative concentration it suffice to prove this for one set D from corollary II.1.6.. Let

$$D = \{(X,Y) \| X=(\frac{1}{N},\ldots,\frac{1}{N}), Y \in \mathbb{R}^N\}.$$

Then

$$C_r|_D(X,Y) = \frac{2}{N-1}[(\frac{1}{N}-\alpha_1)+2(\frac{1}{N}-\alpha_2)+\ldots+N(\frac{1}{N}-\alpha_N)]$$

if $A_Y=(\alpha_i)_{i=1,\ldots,N}$ is arranged decreasingly. It now follows that

$$C_r|_D(X,Y) = \frac{2}{N-1}(\frac{N+1}{2}-\sum_{i=1}^{N} i\alpha_i)$$

$$= C(Y),$$

the classical Pratt measure of Y (cf. Egghe and Rousseau (1990a,b, 1991), Rousseau (1992)). For the other set D we find

$$C_r|_D(X,Y) = C(X). \quad \square$$

As $G = \frac{N-1}{N}C$ (formula (6)) is Gini's index, it is clear that

$$G_r = \frac{2}{N} \max_{\varphi \in \pi_N} \sum_{i=1}^{N} \varphi(i)(a_i-\alpha_i) \qquad (24)$$

is a good measure of symmetric relative concentration too and we also have that, in the notation of proposition II.2.2, $G_r|_D=G$.

In absolute concentration theory, the coefficient of variation

$$V(X) = V = \frac{\sigma}{\mu}, \tag{4}$$

where σ and μ is the standard deviation and the mean of the vector X, is one of the most important concentration measures. We refer to Egghe and Rousseau (1991) for a study of its many good properties. For $X=(x_1,\ldots,x_N)$, the square of (4) reads

$$V^2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}{\left(\frac{\sum_{i=1}^{N}x_i}{N}\right)^2}$$

$$V^2 = N\sum_{i=1}^{N}(a_i-\frac{1}{N})^2 \tag{25}$$

This result gives a clear hint how to construct a possibly good measure of symmetric relative concentration : replace the constant vector $(\frac{1}{N},\ldots,\frac{1}{N})$ by a second one $A_Y=(\alpha_i)_{i=1,\ldots,N}$. We therefore define

$$V_r^2(X,Y) = N\sum_{i=1}^{N}(a_i-\alpha_i)^2. \tag{26}$$

We have the following result

Proposition II.2.3 : V_r^2 (and hence also V_r) is a good measure of symmetric relative concentration.

Proof : It is clear that V_r^2 is symmetric. Suppose now that we have two situations (X,Y) and (X′,Y′) such that

$$A_X - A_Y \text{ ---}< A_{X'} - A_{Y'} \tag{27}$$

and that these vectors are unequal.

The required inequality

$$\sum_{i=1}^{N} (a_i - \alpha_i)^2 < \sum_{i=1}^{N} (a'_i - \alpha'_i)^2 \tag{28}$$

follows from the fact that any function of the form

$$F(Y) = \sum_{k=1}^{N} f(y_k) \tag{29}$$

with f strictly convex satisfies $Y \text{ ---}< Y' \Rightarrow F(Y) \leq F(Y')$; see Hardy, Littlewood and Polya (1988), p. 89. □

We leave it open to construct other good measures of symmetric relative concentration (e.g. based on existing good concentration measures).

Remark : From the above it seems to be a good device - in order to construct new measures of symmetric relative concentration - one starts from a good concentration measure, tries to rewrite it so that it contains factors or terms of the form $a_i - \frac{1}{N}$ in it and then replace $\frac{1}{N}$ by α_i. This is a "rule of thumb" which might be useful but in any case, the final result must still be investigated and a mathematical proof that the measure is a good one (of symmetric relative concentration) must still be given.

We will now study the asymmetric case.

III. Asymmetric relative concentration theory.

III.1. Introduction : Overlap

As in the symmetric case we want to compare two vectors but one of these vectors is fixed, a kind of reference vector. An example could be obtained by considering a fixed database consisting of N fixed parts. Their relative sizes yield this reference vector. Then we can fix an arbitrary topic and study how many documents on this topic are retrieved from each of these parts. Their relative sizes then form the (variable - according to the different topics) second vector.

Because of these conceptual differences we will denote the reference vector of the relative values by $W=(w_1,\ldots,w_N)$ and, as usual, the variable vector by X, resulting in the vector of relative values $A_X=(a_1,\ldots,a_N)$.

By asymmetric relative concentration, we mean the concentration study of the overlap vector of A_X with respect to W. Overlap is a well-known informetric topic but is not so easy to study. However its definition is easy. It originates from overlap of collections (libraries, documentary systems, and so on). Let us consider two collections, symbolized by the sets A and B. Overlap of B with respect to A is the conditional probability of B, given A : $P(B|A)$. Conditional probabilities are calculated as follows :

$$P(B|A) = \frac{P(A\cap B)}{P(A)}, \tag{30}$$

where $P(A\cap B)$ denotes the fraction of documents that are in A and B and where this number is divided by the fraction of documents that are in A. We refer to Fig. 3 for a visualization of this.

Note that overlap is asymmetrical : $P(A|B)\neq P(B|A)$. Fig. 4 illustrates this : here $P(A|B)\approx 0$ and $P(B|A)=1$.

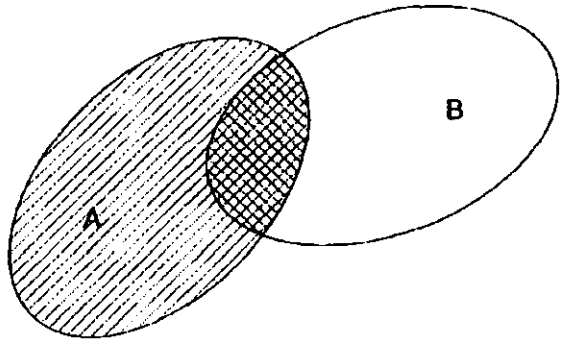

Fig. 3 Overlap of B w.r.t. A

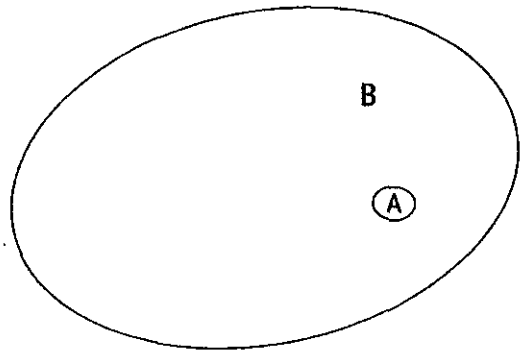

Fig. 4 $P(A|B)\approx 0$ and $P(B|A)=1$

Translated into our framework we hence look, for every "location" i, to $\frac{a_i}{w_i}$, in other words, the tangent of the angle β_i in Fig. 5. We assume that all $w_i > 0$ (it is pointless to consider empty parts when studying overlap !).

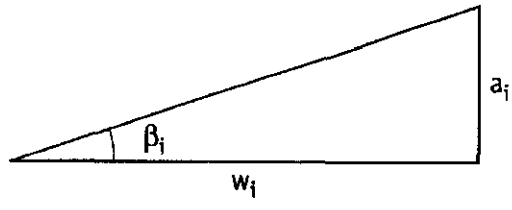

Fig. 5 Graphical illustration of overlap.

It is now clear how asymmetric relative concentration theory must be defined : Fig. 5 is one element in the cumulative weighted Lorenz curve that must be constructed using the normalized weight vector $W=(w_1,\ldots,w_N)$ with $\sum_{i=1}^{N} w_i = 1$.

III.2 Asymmetric relative concentration axiom and properties.

Based on the observations in the previous subsection we have the following definitions.

Definition III.2.1 : The Lorenz curve of a vector X with respect to the normalized weight vector W is the broken line connecting (0,0) with the points

$$\left(\sum_{j=1}^{i} w_j, \sum_{j=1}^{i} a_j \right)_{i=1,\ldots,N} . \tag{31}$$

Here $A_X=(a_1,\ldots,a_N)$ as usual and we have ordered the coordinates such that

$$\frac{a_1}{w_1} \geq \frac{a_2}{w_2} \geq \ldots \geq \frac{a_N}{w_N} , \tag{32}$$

as usual. Such a curve is also called a weighted Lorenz curve.

An illustration of a weighted Lorenz curve is given in Fig. 6. Note that if $W=(\frac{1}{N},\ldots,\frac{1}{N})$ (N coordinates) we obtain the classical Lorenz curves.

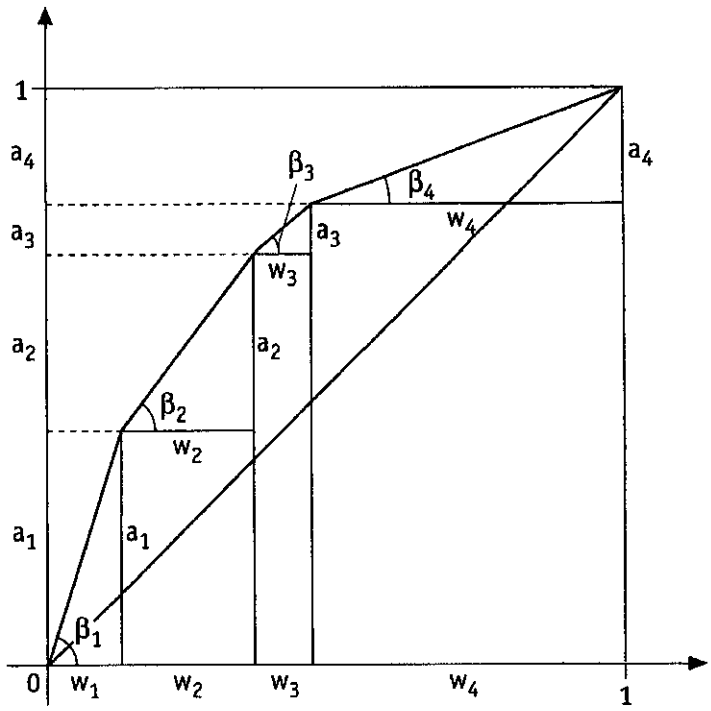

Fig. 6 Illustration of a weighted Lorenz curve.

Let X and X′ be two vectors of N coordinates and denote, as usual, $A_X=(a_1,\ldots,a_N)$, $A_{X'}=(a'_1,\ldots,a'_N)$. The weights $W=(w_1,\ldots,w_N)$ are fixed. We say that X ---< X′ if the weighted Lorenz curve of X is never above the one of X′. We are now in a position to define what are good meausres of asymmetric relative concentration. In view of the above they also measure overlap concentration or overlap inequality.

Definition III.2.2 : Let g_w : $\mathbb{R}^N \to \mathbb{R}$

$X \to g_r(X)$

be any function. We say that g_w is a good measure of asymmetric relative concentration (or a good measure of overlap concentration) if, for any two vectors X, X′ such that X ---< X′ and $X \neq X'$ we have that $g_w(X) < g_w(X')$.

It already follows from this definition that g_w must attain its smallest value in case $A_{\bar{X}}=W$ and its highest value (within a given W) for the vector $A_X=(1,0,\ldots,0)$ where w_1 is the smallest weight.

It is also trivial that, if $W=\left(\frac{1}{N},\ldots,\frac{1}{N}\right)$ (N coordinates) our theory of asymmetric relative concentration reduces to the classical theory of concentration.

Note III.2.3 : It is not the purpose of interchanging the roles of A_X and W. It can even be impossible if some a_is are zero (which is allowed). If we do interchange these roles we are in fact interested in symmetrical results but some measures of asymmetric relative concentration (see further) might give different values.

III.3 Measures of asymmetric relative concentration.

Measures of asymmetric relative concentration are also known as 'weighted' concentration measures (Patil and Taillie (1982), Theil (1967), Rousseau (1992)). If one starts with a concentration measure that satisfies the cell replication axiom (Dalton (1920), Rousseau (1992)) or with the opposite of an eveness measure (Nijssen et al. (1998)) then a rule of thumb to obtain a measure for asymmetric relative concentration is to replace $\frac{1}{N}$ by w_i. This leads to the following measures that all respect the partial order imposed by weighted Lorenz curves.

1. The asymmetric (or weighted) Theil measure :

$$Th_w(X) = \sum_{i=1}^{N} a_i \ln\left(\frac{a_i}{w_i}\right) \tag{33}$$

2. The asymmetric (or weighted) squared coefficient of variation :

$$V_w^2(X) = \sum_{i=1}^{N} \frac{(a_i-w_i)^2}{w_i} \tag{34}$$

Another form of the weighted squared coefficient of variation is :

$$V_w^2(X) = \sum_{i=1}^{N} \frac{1}{w_i}a_i^2-1 \tag{35}$$

3. The asymmetric (or weighted) Gini index :

$$G_w(X) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} |w_i a_j - w_j a_i| \qquad (36)$$

Its meaning is the same as that of the (unweighted) Gini index, namely twice the area between the Lorenz curve and the diagonal. This index has been used in studies of the localization of industry, under the name of 'locational Gini coefficients' (Krugman (1991)).

Note that, if we compare V_w^2 (formula (34)) with V_r^2 (formula (26)), we see that V_r^2 is a global measure while V_w^2 is an average measure, in the sense of Egghe and Rousseau (1996). Indeed (34) rewrites as

$$V_w^2 = N\left(\frac{1}{N} \sum_{i=1}^{N} \frac{(a_i - w_i)^2}{w_i} \right)$$

while (26) rewrites (since $\sum_{i=1}^{N} w_i = 1$)

$$V_r^2 = N\left(\frac{\sum_{i=1}^{N} (a_i - w_i)^2}{\sum_{i=1}^{N} w_i} \right)$$

where we replace here α_i by w_i for comparison reasons. From Egghe and Rousseau (1996) we know that V_w^2 can be $>$, $<$ or $=$ V_r^2 according to the sign of the slope r of the regression line of the cloud of points

$$\left(w_i, \frac{(a_i - w_i)^2}{w_i} \right)_{i=1,\ldots,N}$$

We have $V_w^2 > V_r^2 \Leftrightarrow r < 0$

$V_w^2 < V_r^2 \Leftrightarrow r > 0$

$V_w^2 = V_r^2 \Leftrightarrow r = 0$.

This difference between V_w and V_r clearly underlines the difference between symmetric and asymmetric theory of relative concentration.

That the above measures are good measures of asymmetric relative concentration is implicite in Theil (1967) and Patil and Taillie (1982) but their proofs are unclear. We present now a new and self-contained proof.

Let g_w denote any of the measures Th_w, V_w^2 (V_w) or G_w. We first prove a lemma and a consequence of this lemma. Then we will state and prove the major result.

Lemma III.3.1 : Let $g_w(X)$ be the value of g_w in $X=(x_1,\ldots,x_N)$, where the fixed weight vector is given by $W=(w_1,\ldots,w_N)$, where $w_j \in Q^+$ (the positive rational numbers), $j=1,\ldots,N$. Denote $w_j=\frac{n_j}{d_j}$, n_j, $d_j \in \mathbb{N}$ for all $j=1,\ldots,N$. Then $g_w(X)$ is equal to the unweighted value of g_w for $A_X=(a_1,\ldots,a_N)$ replaced by the vector consisting of F_j times $\frac{a_j}{F_j}$, where $F_j=Pw_j$ $(j=1,\ldots,N)$ and $P=\prod_{j=1}^{N} d_j$.

Proof : Note first that the latter vector has length P. We will prove the assertion for the measures Th_w, V_w^2 (hence V_w) and G_w separately.

(i) $g_w = Th_w$.

Th_w in the unweighted vector

$$\underbrace{\frac{a_1}{F_1}, \cdots, \frac{a_1}{F_1}}_{F_1 \text{ times}}, \cdots, \underbrace{\frac{a_N}{F_N}, \cdots, \frac{a_N}{F_N}}_{F_N \text{ times}}$$

is

$$\ln P + \sum_{i=1}^{N} F_i \frac{a_i}{F_i} \ln\left(\frac{a_i}{F_i}\right)$$

$$= \quad \ln P + \sum_{i=1}^{N} a_i \ln a_i - \sum_{i=1}^{N} a_i \ln P - \sum_{i=1}^{N} a_i \ln w_i$$

$$= \quad \sum_{i=1}^{N} a_i \ln\left(\frac{a_i}{w_i}\right) = Th_w(X), \text{ by (33).}$$

(ii) $g_w = V_w^2$

V_w^2 in the unweighted vector above equals

$$P \sum_{i=1}^{N} F_i \left(\frac{a_i}{F_i}\right)^2 - 1$$

$$= \quad \sum_{i=1}^{N} P \frac{1}{F_i} a_i^2 - 1 = \sum_{i=1}^{N} \frac{a_i^2}{w_i} - 1 = V_w^2(X), \text{ by (35).}$$

(iii) $g_w = G_w$

G_w in the unweighted vector above equals

$$\frac{1}{2P} \sum_{k=1}^{N} \sum_{l=1}^{N} F_k F_l \left|\frac{a_k}{F_k} - \frac{a_l}{F_l}\right|$$

$$= \quad \frac{1}{2P} \sum_{k=1}^{N} \sum_{l=1}^{N} |F_l a_k - F_k a_l|$$

$$= \quad \frac{1}{2P} \sum_{k=1}^{N} \sum_{l=1}^{N} |w_l a_k - w_k a_l| = G_w(X), \text{ by (36).} \quad \square$$

Corollary III.3.2 : Let X, W and Y, W′ be two weighted systems yielding the same Lorenz curves and where the coordinates in W and W′ are rational. Then $g_w(X) = g_w(Y)$.

Proof :The proof follows the lines of the above lemna. Another argument is simply applying the above lemna twice and noting that the unweighted version of the measures g_w are only dependent on the Lorenz curves and not on the number N of divisions of [0,1]. □

Proposition III.3.3 : The measures Th_w, V_w^2 (hence V_w) and G_w are good measures of asymmetric relative concentration.

Proof : That the above measures (commonly denoted by g_w) are good measures for rational weight vectors W is clear from the above lemma and the fact that the unweighted versions of g_w are good measures of concentration (see e.g. Egghe and Rousseau (1990a,b), Rousseau (1992)). Let now $X=(x_1,\ldots,x_N)$ $---<$ $X'=(x'_1,\ldots,x'_N)$, $X\neq X'$, where $---<$ is defined via $W=(w_1,\ldots,w_N)$, where $w_i\in\mathbb{R}^+$ for all $i=1,\ldots,N$.

It is then easy to see that vectors $Y\neq Y'$ exist and a weight vector $W'=(w'_1,\ldots,w'_N)$, where $w'_i\in Q^+$ for $i=1,..,N$ such that Y $---<$ Y' (here $---<$ is defined via W') and such that the Lorenz curve of Y (via W') is above the one of X (via W) and the Lorenz curve of X' (via W) is above the one of Y' (via W'). Let us denote this, shortly, by X $---<$ Y $---<$ Y' $---<$ X'. Since g_w is good for rational weight vectors, we have that $g_w(Y) < g_w(Y')$. Let $\delta=g_w(Y')-g_w(Y)>0$. Note that the considered functions above are all continuous on the space of all Lorenz curves with distance function (between two Lorenz curves) the maximum of the distances between the corresponding points with the same abscissa. It is hence easy to construct vectors Z, Z' such that Z $---<$ Y $---<$ Y' $---<$ Z' and such that $|g_w(Z)-g_w(X)| < \frac{\delta}{4}$ and $|g_w(Z')-g_w(X')| < \frac{\delta}{4}$.

Here all $---<$ are with respect to an existing $W''=(w''_1,\ldots,w''_M)$, $w''_j\in Q^+$ for all $j=1,\ldots,M$ where W'' refines W', i.e. the Lorenz curves w.r.t. W' are the same as w.r.t. W'' but the latter one is "cut" into more pieces. Note that the value of g_w is only dependent on the Lorenz curve and not on the number of rational pieces, as follows from corollary III.3.2. Hence, the values of $g_w(Y)$ and $g_w(Y')$ are unchanged when going from W' to W''.

Consequently, we now have :

$$g_w(X') - g_w(X)$$

$$= g_w(X')-g_w(Z')+g_w(Z')-g_w(Y')+g_w(Y')-g_w(Y)+g_w(Y)-g_w(Z)+g_w(Z)-g_w(X)$$

$$> g_w(Y') - g_w(Y) - \frac{\delta}{4} - \frac{\delta}{4} = \frac{\delta}{2} > 0$$

since $g_w(Y') < g_w(Z')$ and $g_w(Y) > g_w(Z)$, again since g_w is good for rational weight vectors.

This shows that the measures g_w are good for all weight vectors. It also follows that $g_w(X)$ is only dependent on the weighted Lorenz curve and not on the weight vector W itself. □

IV. Applications

IV.1 Indexing and information retrieval.

A very classical model of indexation and of information retrieval (IR) is the vector space model, see e.g. Salton and Mc Gill (1987) or Egghe and Rousseau (1998). Here we have a document space DS and a query space QS both being subsets of $[0,1]^N$. Here N is the total number of key words used in the system. A document $d \in DS$ or a query $q \in QS$ is then of the form $(x_1,\ldots,x_N)$ where x_i denotes the degree of importance of key word i in d or q. Most classical $x_i=0$ (key word i does not appear in d or q) or $x_i=1$ (key word is appearing in d or q).

The degree of similarity between a document $d=(d_1,\ldots,d_N)$ and a query $q=(q_1,\ldots,q_N)$ can be described via the cosine of the angle between d and q :

$$\cos(d \hat{} q) = \frac{\sum_{i=1}^{N} d_i q_i}{\sqrt{\sum_{i=1}^{N} d_i^2}\sqrt{\sum_{i=1}^{N} q_i^2}} \qquad (37)$$

(see the references given above for an explanation). By taking

$$1 - \cos(d\hat{}q) \tag{38}$$

we express dissimilarity between d and q (note that since the angle between d and q is between 0 and $\frac{\pi}{2}$ radians, $\cos(d\hat{}q) \geq 0$). Studying (38) is equivalent to studying (37), so we will study (38) since in this paper everything has been expressed in the context of concentration.

(37) is a very famous measure of similarity between a document and a query, so (38) is expected to be a good measure of relative concentration between d and q. It is clear that (38) is symmetric. However we can prove the following result.

Proposition IV.1.1 : The function

$$(d,q) \rightarrow 1 - \cos(d\hat{}q)$$

is not a good measure of symmetric relative concentration.

Proof : The proof is essentially contained in Egghe (1990) (section II.2) but in other terminology. Therefore we give a direct proof here. Proving that $1 - \cos(d\hat{}q)$ is not a good measure of symmetric relative concentration only requires a counterexample. A whole set of counterexamples is given as follows. Let

$$d = (\overbrace{1,\ldots,1}^{M_1},0,\ldots,0)$$

such that $M_1 < \frac{N}{2}$

$$q = (0,\ldots,0,\overbrace{1,\ldots,1}^{M_2})$$

such that $M_2 < \frac{N}{2}$

$$d' = (\underbrace{1,\ldots,1}_{M_1'},0,\ldots,0)$$

such that $M_1' < M_1$, and

$$q' = (0,\ldots,0,\underbrace{1,\ldots,1}_{M_2'})$$

such that $M_2' < M_2$. We assume that all vectors have length N.

Then it is easy to see that

$$A_d - A_q = \left(\underbrace{\frac{1}{M_1},\ldots,\frac{1}{M_1}}_{M_1},0,\ldots,0,\underbrace{-\frac{1}{M_2},\ldots,-\frac{1}{M_2}}_{M_2}\right)$$

and

$$A_{d'} - A_{q'} = \left(\frac{1}{M_1'},\ldots,\frac{1}{M_1'},0,\ldots,0,-\frac{1}{M_2'},\ldots,-\frac{1}{M_2'}\right)$$

where there are more 0s in $A_{d'}$-$A_{q'}$ than in A_d-A_q (but also in this vector, 0 occurs).

The Lorenz curve of A_d-A_q has the following values in the abscissae $\frac{1}{N},\frac{2}{N},\ldots,1$:

$$\frac{1}{M_1},\frac{2}{M_1},\ldots,1,1,\ldots,1,1-\frac{1}{M_2},1-\frac{2}{M_2},\ldots,0 \qquad (39)$$

The one of $A_{d'}$-$A_{q'}$ has the following values in the same abscissae

$$\frac{1}{M_1'},\frac{2}{M_1'},\ldots,1,1,\ldots,1,1-\frac{1}{M_2'},1-\frac{2}{M_2'},\ldots,0 \qquad (40)$$

and the set of abscissae where we have 1 in (39) is a strict subset of the set of abscissae where we have 1 in (40). Since $M_1' < M_1$ and $M_2' < M_2$ we see that the Lorenz curve of $A_{d'}$-$A_{q'}$ is always strictly above the one of A_d-A_q (except in (0,0) and (1,0) where they coincide of course). This is readily seen for the sequence

$$\frac{1}{M_1'},\frac{2}{M_1'},\ldots,1,1,\ldots,1$$

in (40). At the last 1 in (40), the corresponding value in (39) is of the form $1-\frac{k}{M_2}<1$ and from the next coordinate on both curves then decrease to zero in a linear way, showing that indeed

$$A_d - A_q \text{ ---} < A_{d'} - A_{q'} \tag{41}$$

Furthermore A_d-$A_q \neq A_{d'}$-$A_{q'}$. Hence, in order to be a good measure of symmetric relative concentration we must have that

$$1 - \cos(d\hat{}q) < 1 - \cos(d'\hat{}q') \tag{42}$$

But it is trivial to see, since M_1, M_2, $M_1', M_2' < \frac{N}{2}$ that $\cos(d\hat{}q) = \cos(d'\hat{}q') = 0$, contradicting (42). □

The intuitive idea behind this is that $\cos(d\hat{}q)$ is insensitive for the degree of broadness of a document (or a query). Let us illustrate this by giving some examples (see also Egghe (1990)).

1. d = (1,0,...,0), q = (0,...,0,1)
2. d = (1,...,1,0), q = (0,...,0,1)
3. d = (1,...,1,0,...,0), q = (0,...,0,1,...,1)

   where the coordinate of the last 1 in d is smaller than the coordinate of the first 1 in q.

It is trivial to see that in all these cases, $\cos(d\hat{}q)=0$ while d in example 1 is very specialized in a wrong topic for q but d in example 2 is a document with a very broad scope. So, from IR point of view, although no d matches no q very well, we think that d of example 2 has "some" (low) interest w.r.t. q, certainly more than d of example 1. Example 3 represents a more "real life" example. In a ranked output we would like to see d of example 2 to have a smaller rank (i.e. ranked earlier) than d of example 1 and with d of example 3 somewhere between them. This is important when thresholds apply to cut off a list of documents (e.g. provided by a search engine in WWW). Formulae (37) or (38) are not capable of doing this as proved in proposition IV.1.1.

We know already that all good measures of symmetric relative concentration (e.g. C_r, V_r) will yield such "fine tuned" rankings. In Egghe (1990) we examined the similarity measure $D_r=1-\frac{C_r}{2}$ (since $0\leq C_r\leq 2$, we have that $0\leq D_r\leq 1$)

Proposition IV.1.2 (Egghe). Let

$$d = (\underbrace{1,\ldots,1}_{M_1},0,\ldots,0) \text{ and}$$

$$q = (0,\ldots,0,\underbrace{1,\ldots,1}_{M_2})$$

such that $M_1 \leq N-M_2$, then

$$D_r = 1 - \frac{1}{2}\frac{N(N-M_1)-(N-M_2)(N-M_1-M_2)}{(N-1)M_2} \qquad (43)$$

$$> 0 .$$

Hence D_r is sensitive with respect to changes in M_1 and M_2, contrary to $1 - \cos(d\hat{}q)$. Of course also V_r could be used here (or rather $1-\frac{V_r}{2}$). We leave it as an open problem (and a challenge !) to investigate the retrieval power of these new measures.

In this subsection we have restricted our attention to the study of the symmetric relative concentration of (d,q). In exactly the same way we can study the symmetric relative concentration (or rather similarity) of two documents (d_1,d_2) or two queries (q_1,q_2). Especially the similarity between two documents (d_1,d_2) is interesting for indexing purposes.

IV.2 Content locality.

In Viles and French (1999), the authors want to study content locality in distributed documentary systems, i.e. a set of autonomous distinct document collections (sites). These sites form a reference frame with which comparisons are made. This goes as follows. For an arbitrary topic one checks how many documents on this topic exist in the different sites. Then one compares these relative scores with the relative sizes of these sites.

When the topic is evenly distributed over the sites, both vectors of relative numbers are equal. Content locality wants to measure the opposite : how "different" is the vector of relative number of documents in each site from the vector of relative site sizes ?

In other words the latter vector is the fixed reference vector with which each topic is compared. Again in other words, one is interested in the overlap of the topic in the different sites. From this it directly follows that Viles and French want to compare these vectors from the asymmetric relative concentration point of view. Instead they used a (admittingly good) measure of symmetric relative concentration, namely a measure proportional to V_r. It is indeed easy to see that their measure σ_t (Viles and French (1999), p. 321) is nothing else than $\frac{V_r}{\sqrt{N}}$ as given by formula (26) here. Hence the inequality in topical overlap is not measured in the exact way. Our advise is e.g. to use any of the measures (33)-(36), being good measures of asymmetric relative concentration (and since they used $\frac{V_r}{\sqrt{N}}$ we even advise to use V_w or $\frac{V_w}{\sqrt{N}}$). Note that in section III.3 we already remarked the difference between V_r and V_w, underlining the different nature of these measures and the fact that one measure cannot be used as a substitute for the other. This is another argument to use V_w and not V_r in this connection.

Finally we want to make a few remarks on the arguments given by Viles and French (1999) concerning the quality of their measures σ_t. First of all we admit that σ_t is a good measure of symmetric relative concentration. Hence the qualities are simply given by definition II.1.1 ; nothing else is needed. The two properties given in Viles and French (1999) discussed on p. 321 are not correct properties and, in addition, they are only verified by Viles and French in the very special case that all sizes are equal. In our frame work this means concentration and hence the results of concentration theory apply (Egghe and Rousseau (1990 a,b, 1991), Rousseau (1992)), which are much more general than these two properties.

With this in mind let us discuss these two properties (the text in quotes (") is the assertion of Viles and French)

1. "As fewer (more) sites contain members of some topic t, measured locality should increase (decrease)." As said this is only true in case all sites have the same size. When this is not the case, measured locality can be very high even if topic t appears in every site. This happens if t has (relatively) the most documents in the smallest sites. This property is covered by the general definition II.1.1.
2. "Given that k sites contain members of t, the more asymmetric the distribution of these members, the higher measured locality should be". Again this is only true in case all sites have the same size. Assertion 2 only talks about the distribution of the topic vector. It is obvious that - for general site sizes - we must make a relative comparison between the topic vector and the site vector. Note the term "asymmetric" in their assertion 2.

In short, the only correct way of describing content locality is by applying one of the existing good measures of asymmetric relative concentration.

V. Conclusions and open problems.

We have extended the theory of concentration of a vector to the case of relative concentration of one vector to another. We found that, for practical applications, there is a

need for (at least) two models of relative concentration : symmetric and asymmetric relative concentration.

In the first model the measures f_r are symmetric : $f_r(X,Y)=f_r(Y,X)$ for all vectors X and Y. The basic requirement for f_r to be a good measure of symmetric relative concentration is that

$$A_X - A_Y \text{ ---}< A_{X'} - A_{Y'}$$

(and not equal) implies

$$f(X,Y) < f(X',Y') .$$

Concrete measures are given : the relative Pratt measure C_r and the relative variation coefficient V_r. It is proved that the cosine matching function has major drawbacks and hence cannot be used to measure symmetric relative concentration.

In the second model a variable vector X is compared with a fixed reference vector W. In this theory, measures g_w are good measures of asymmetric concentration if

$$X \text{ ---}< X'$$

and $X \neq X'$ imply $g_w(X) < g_w(X')$. Here ---< denotes the weighted Lorenz order (weighted by the reference vector W). Concrete measures are given and we show that content locality (of Viles and French) is nothing else than measuring asymmetric relative concentration of a topic vector with the vector of relative sizes of the sites in documentary systems.

As an open problem we can ask for the construction of other good measures of symmetric or asymmetric relative concentration. The rules of thumb to do so (given in the respective sections) should be examined. The Viles and French calculations with their σ_t should be redone, now using any good measure of asymmetric relative concentration. In addition,

from the IR side, performance analyses should be executed using different good measures of symmetric relative concentration, instead of using Salton's cosine formula.

References

M.P. Carpenter (1979). Similarity of Pratt's measure of class concentration to the Gini index. Journal of the American Society for Information Science 30, 108-110.

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

L. Egghe (1988). The relative concentration of a journal with respect to a subject and the use of online services in calculating it. Journal of the American Society for Information Science 39(4), 281-284.

L. Egghe (1990). A new method for information retrieval based on the theory of relative concentration. Proceedings of the 13th international Conference on Research and Development in Information Retrieval (SIGIR), Brussels, 469-493.

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

L. Egghe and R. Rousseau (1990b). Elements of concentration theory. Informetrics 89/90 (L. Egghe and R. Rousseau, eds.), 97-137.

L. Egghe and R. Rousseau (1991). Transfer principles and a classification of concentration measures. Journal of the American Society for Information Science 42(7), 479-489.

L. Egghe and R. Rousseau (1996). Average and global impact of a set of journals. Scientometrics 36(1), 97-107.

L. Egghe and R. Rousseau (1997). Duality in information retrieval and the hypergeometric distribution. Journal of Documentation 53(5), 488-496.

L. Egghe and R. Rousseau (1998). Topological aspects of information retrieval. Journal of the American Society for Information Science 49(13), 1144-1160.

C. Gini (1909). Il diverso accrescimento delle classi sociali e la concentrazione della ricchezza. Giornale degli Economisti, serie 11, 37.

G. Hardy, J.E. Littlewood and G. Polya (1988). Inequalities. Cambridge University Press, Cambridge.

P. Krugman (1991). Geography and Trade. University Press, Leuven.

D. Nijssen, R. Rousseau and P. Van Hecke (1998). The Lorenz curve : a graphical representation of evenness. Coenoses 13, 33-38.

G.P. Patil and C. Taillie (1982). Diversity as a concept and its measurement. Journal of the American Statistical Association 77, 548-561.

A.D. Pratt (1977). A measure of class concentration in bibliometrics. Journal of the American Society for Information Science 28, 285-292.

R. Rousseau (1992). Concentration and diversity in informetric research. Ph. D. Thesis, University of Antwerp (UIA).

G. Salton and M.J. Mc Gill (1987). Introduction to modern Information Retrieval. Mc Graw-Hill, Singapore.

H. Theil (1967). Economics and Information Theory. North-Holland, Amsterdam.

C.L. Viles and J.C. French (1999). Content locality in distributed digital libraries. Information Processing and Management 35, 317-336.