Properties of the n-overlap vector and n-overlap similarity theory

Properties of the n-overlap vector and n-overlap similarity theory

by

L. Egghe

Universiteit Hasselt (UHasselt), Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

leo.egghe@uhasselt.be

ABSTRACT

In the first part of this paper we define the n-overlap vector whose coordinates consist of the fraction of the objects (e.g. books, N-grams,…) that belong to 1, 2,…, n sets (more generally: families) (e.g. libraries, databases,…). With the aid of the Lorenz concentration theory we build a theory of n-overlap similarity and corresponding measures, such as the generalized Jaccard index (generalizing the well-known Jaccard index in case $n = 2$).

Next we determine the distributional form of the n-overlap vector assuming certain distributions of the object's and of the set (family)-sizes. In this section the decreasing power law and decreasing exponential distribution is explained for the n-overlap vector. Both item (token) n-overlap and source (type) n-overlap are studied.

---

[1] Permanent address

Key words and phrases: n-overlap vector, Lorenz, Jaccard index, power law, N-gram

The final section is devoted to the n-overlap properties of objects indexed by a hierarchical system (e.g. books indexed by numbers from a UDC or Dewey system or by N-grams). We show that the results of Section II can be applied here. We also show that the Lorenz-order of the n-overlap vector is respected by an increase or a decrease of the level of refinement in the hierarchical system (e.g. the value N in N-grams).

I. Introduction

Overlap is an important topic in information science. The simplest type of overlap is the one between two sets A and B, where A and B can be libraries, databases or (more generally) collections of objects and where one counts the number of objects that are common to both sets. So, essentially, one studies $|A \cap B|$, the number of objects in the intersection of A and B. Next one determines relative measures of overlap, where one divides $|A \cap B|$ by a "normalizing" number e.g.

$$J = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

(Jaccard's index) or the asymmetric measures

$$O_1 = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{2}$$

$$O_2 = \frac{|A \cap B|}{\max(|A|, |B|)} \tag{3}$$

(cf. also Salton and McGill (1983), Gluck (1990), Hood and Wilson (2003), Egghe and Rousseau (2005)). Other measures can be built from (2) and (3), e.g. the harmonic average of O_1 and O_2 (leading to Dice's coefficient) or the geometric mean of O_1 and O_2 (leading to Salton's cosine measure) – see again Salton and McGill (1983) or Egghe and Rousseau (2005).

Basic in overlap is the notion of equality, required for an exact definition of the sets A and B. Mathematically this is a clear notion (boiling down to the concept of element of a set) but in practise (in information science) it is a delicate notion to define. Let us consider the case of two libraries A and B. How can their overlap be defined? Everything depends on what one wants to study. From the cataloger's point of view (and in the connection of making union catalogs) one uses the finest methods to make a distinction between books e.g. expressed by the ISBN. Here soft- and hardcovers of the "same" book are considered to be different. For the library user, most probably, they will be considered the same. Even old and new editions of the "same" book can be considered the same by library users if their content is the same (or almost the same).

In a way we can say that overlap is determined by the hierarchical level of the description (indexing) of an object and this object's description replaces the object in the sets A and B. An example is the description of books in libraries using a topical numbering systems such as UDC or Dewey or the description of articles using subject-dependent classification systems. Another example is given by using N-grams for the book description (e.g. the first N letters in the title or a combined system of letters in authors' names and title's words - see e.g. Egghe (2000) and references therein). It is clear that the degree of refinement (expressed by the length of the UDC or Dewey number or by N, the number of letters (more generally: symbols) in an N-gram) of the indexing is of direct influence on the measurement of the overlap of sets A and B.

A second important aspect in the definition of overlap is the fact that overlap can be measured using type or using token (in the linguistical terminology). Type means the object as such and token the uses (or occurrences) of this object. So does it make a difference whether or not we consider a book, of which 5 identical copies are in a library, as being once (type) or 5 times (token) in a library and the same for the overlap between two libraries (also in the connection of relative measures of overlap: do we consider the total number of token or of type ?). Both types of overlap will be studied in this article. Since sets can only contain an element once it is better to consider A and B as families, the mathematical notion of "sets" where identical elements can occur.

Notationally we can also consider sets A and B as consisting of objects and where we consider their (hierarchical) description (i.e. indexing) as type, denoted by $F(A)$ and $F(B)$ or as token, denoted by $G(A)$ and $G(B)$. Let us give a simple example. A and B are libraries. Let A have 6 books and let B have 4 books. Let their 3-gram descriptions be (as token) (with index 3 to denote 3-grams)

$$G_3(A) = \{abc, abb, abc, cad, dbc, ccc\}$$
$$G_3(B) = \{abc, abb, abb, abc\}$$

then we have

$$F_3(A) = \{abc, abb, cad, dbc, ccc\}$$
$$F_3(B) = \{abc, abb\}$$

If we consider their 2-gram descriptions then we have (right truncation)

$$G_2(A) = \{ab, ab, ab, ca, db, cc\}$$
$$G_2(B) = \{ab, ab, ab, ab\}$$

and

$$F_2(A) = \{ab, ca, db, cc\}$$
$$F_2(B) = \{ab\}$$

Note also that $|G_3(A)| = |G_2(A)|$ and the same for B and that this equality is false for F.

Particularly when we study N-gram overlap (in the last section), the above notation will be handy.

So, summarizing, the study of overlap (between two families A and B) requires the definition of equality of objects (which we consider to be established, throughout this paper) and the choice between type overlap or token overlap (exact definitions will be given in Section II – in Section I it does not matter whether we consider the overlap for tokens or for types).

Let us now turn our attention to n-overlap meaning aspects of overlap between n families, denoted as $D_1,...,D_n$ (recall that we keep the full generality by using families but these can be ordinary sets as well, of course). Essentially we can consider n aspects of overlap namely objects belonging to only one family or to exactly 2 families, …, or to all n families. Let us denote, for $k = 1,...,n$

$$\varphi(k) = \text{fraction, with respect to the union of all families, of the objects that belong to exactly k families} \tag{4}$$

(we will specify this more in Sections II and III also making the difference between type and token overlap).

Then we can consider the n-overlap vector

$$\varphi = (\varphi(1),\varphi(2),...,\varphi(n)) \tag{5}$$

Note that only $n-1$ of these n coordinates are independent since

$$\sum_{k=1}^{n} \varphi(k) = 1. \tag{6}$$

The study of this vector will be the main topic of this paper. Theoretically, φ can be any vector satisfying (6). There are not many studies that deal with n-overlap. In Hood and Wilson (2003) one presents a 12-overlap vector (in table form) on the overlap of records on fuzzy set theory across 12 databases. More correctly, the n in Hood and Wilson, is not 12 but (as they indicate) "over 100" where all values $\varphi(k) = 0$ for $k > 12$. In this sense, the data of Hood

and Wilson (2003) fit into our framework (5). Their twelve $\varphi(1),...,\varphi(12)$ values are as in table 1 (multiplied by 15,644 to yield the actual numbers).

Table 1. Hood and Wilson (2003) data

i	$15{,}644\,\varphi(i)$
1	9,897
2	1,922
3	1,299
4	1,209
5	781
6	344
7	122
8	40
9	8
10	10
11	7
12	5
Total	15,644

As is clear from Table 1, $\varphi(k)$ is decreasing (except for 2 objects in $\varphi(10)$). It is indeed a general intuition that φ should be decreasing. An elementary (but too elementary) explanation of this is as follows: let p be the probability for an object to belong to one database ($0 < p < 1$ and assumed to be equal for all databases). If independence applies (which is not always the case) then we have that the probability to belong to k databases is p^k, hence $\varphi(k) = p^k$, a decreasing function.

An example of the most extreme case of an increasing vector φ, being $\varphi = (0,0,...,0,1)$ is given by n databases of documents where their 1-gram indexing (i.e. rough indexing with one letter) comprises all 1-grams $A,...,Z$ in all databases, hence $\varphi(1) = \varphi(2) = ... = \varphi(n-1) = 0$ and $\varphi(n) = 1$.

In the next section we will study the n-vector φ from a Lorenz-curve point of view. Here we suppose φ to be decreasing or increasing. We will determine the Lorenz-curve of φ and we

deduce, in case $n = 2$, the classical overlap similarity measure J (formula (1)) from the Lorenz-curve. Next we generalise this method for general n and present J_n, the n-overlap similarity measure, extending $J_2 = J$. To the best of our knowledge it is the first time that a similarity measure for n-overlap is presented and the presented Lorenz theory of n-overlap allows for comparisons of n-overlap vectors.

In the third section we will derive explicite functional forms for the n-overlap vector φ for type as well as token n-overlap which we will define in an exact way. Based on well-established type-token distributions (e.g. Zipfian ones) and on well-established rank-size distributions for databases (e.g. Zipfian or decreasing exponential) we derive the decreasing power-law functionality or the decreasing exponential functionality for the vector φ, the most evident candidates for a decreasing n-overlap vector and fitted in Hood and Wilson (2003) on the data of Table 1. From these results we derive that the Lorenz-concentration of the vector φ increases with increasing Lorenz-concentration of the type-token distribution or of the rank-size distribution of the databases. We also prove that in these cases the Lorenz-concentration of the type n-overlap vector is always higher than the one of the token n-overlap vector.

The last section studies n-overlap from a hierarchical point of view. Here only token n-overlap can be treated. We prove that, for decreasing n-overlap vectors φ, the Lorenz-concentration curve increases with increasing hierarchical refinement. A similar result is proved for increasing n-overlap vectors. An example of this is given by N-grams with increasing N, illustrating increasing hierarchical refinement.

II. Lorenz theory of the n-overlap vector and a similarity measure of Jaccard-type for n-overlap

In this section any type of overlap can be considered (type or token): we just suppose we have an overlap vector $\varphi = (\varphi(1),...,\varphi(n))$ as in (5) restricted to (6). Hence each $\varphi(i)$ $(i = 1,...,n)$

represents the fraction, with respect to the union of all families, of the objects that belong to exactly i families from the system of families $D_1,...,D_n$.

The case $n = 2$ will show us how to treat the general case. Here we have that $\varphi = (\varphi(1),\varphi(2))$, with

$$\varphi(1)+ \varphi(2)= 1 \tag{7}$$

The number $\varphi(2)$ is the fraction of objects that belong to both families D_1 and D_2. Hence, by (1)

$$\varphi(2)= J \tag{8}$$

in case $n = 2$ and in case $A = D_1$, $B = D_2$ are sets. Note that the above is already a generalization of the Jaccard index in case D_1, D_2 are families and where we consider token 2-overlap: $\varphi(2)$ equals the fraction of the tokens that appear in both families D_1, D_2.

Let us suppose that φ is decreasing (increasing is also possible, covering all cases if $n = 2$; the Lorenz-curves in this case are mirrors (over the diagonal $y = x$) of the ones obtained in case φ decreases). So $\varphi(1)^3 \varphi(2)$. The Lorenz-curve of φ consists of the straight line connection of $(0,0)$ with $\left(\frac{1}{2},\varphi(1)\right)$ with $(1,1)$ (see Egghe and Rousseau (2001) for a simple description of how to construct Lorenz-curves or see the general description (general n) below). Its graph L_φ can be depicted as in Fig. 1.

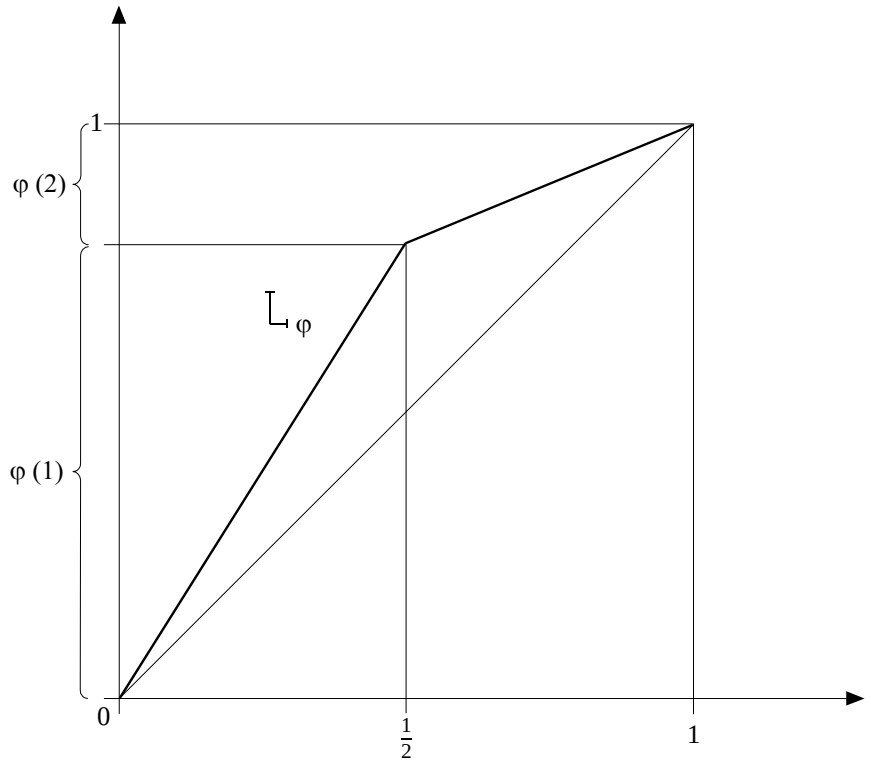

Fig. 1. Lorenz-curve of φ for $n = 2$.

The Lorenz-theory gives that, the higher L_φ, the more concentrated (unequal) are the values $(\varphi(1), \varphi(2))$. The area under L_φ (related to the well-known Gini index – cf. Egghe and Rousseau (1990)) is known to be a good measure of concentration because it gives higher values for more concentrated (i.e. higher Lorenz-curves L_φ) situations. This area θ equals

$$\theta = \frac{3}{4}\varphi(1) + \frac{1}{4}\varphi(2)$$

$$\theta = \frac{3}{4} - \frac{J}{2} \qquad (9)$$

using (7) and (8). Hence

$$J = \frac{3}{2} - 2\theta \tag{10}$$

is a good measure of the opposite of inequality in the vector φ, hence is a good measure of similarity in the vector φ hence of 2-overlap similarity. In general, any decreasing function of θ can be considered as a good measure of similarity. We will now extend these ideas to general n-overlap vectors $\varphi = (\varphi(1),...,\varphi(n))$.

We suppose φ to be decreasing. The Lorenz-curve of φ connects (linearly) $(0,0)$ with the points

$$\left(\frac{i}{N}, \sum_{j=1}^{i} \varphi(j)\right) \tag{11}$$

$i = 1,...,n$, this point being $(1,1)$ for $i = n$, by (6). Its graph is depicted in Fig. 2.

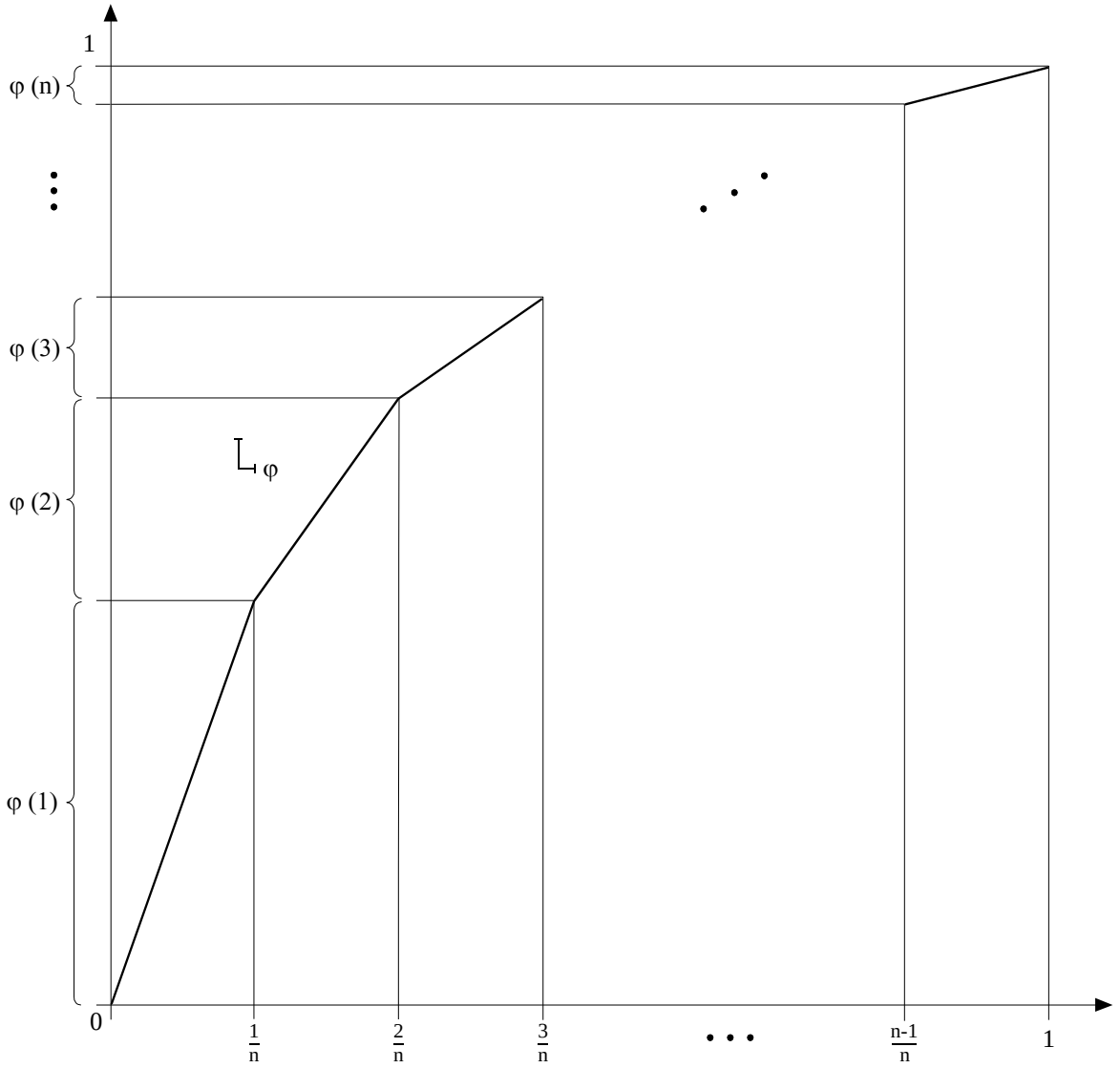

Fig.2. Lorenz-curve of φ, general n

Let us now calculate (as we did in the case $n = 2$) the value θ being the area under L_φ:

$$\theta = \frac{1}{2}\frac{\varphi(1)}{n} + \frac{\varphi(1)}{n} + \frac{1}{2}\frac{\varphi(2)}{n} + \frac{1}{n}(\varphi(1) + \varphi(2)) + \frac{1}{2}\frac{\varphi(3)}{n}$$

$$+ \frac{1}{n}(\varphi(1) + \varphi(2) + \varphi(3)) + \frac{1}{2}\frac{\varphi(4)}{n} + \ldots + \frac{1}{n}(\varphi(1) + \ldots + \varphi(n-1))$$

$$+ \frac{1}{2}\frac{\varphi(n)}{n}$$

$$\theta = \varphi(1)\left(\frac{1}{2n} + \frac{n-1}{n}\right) + \varphi(2)\left(\frac{1}{2n} + \frac{n-2}{n}\right) + \varphi(3)\left(\frac{1}{2n} + \frac{n-3}{n}\right)$$

$$+ \ldots + \varphi(n-1)\left(\frac{1}{2n} + \frac{1}{n}\right) + \varphi(n)\frac{1}{2n}$$

$$\theta = \sum_{k=1}^{n} \varphi(k)\frac{2n-(2k-1)}{2n}$$

$$\theta = 1 - \sum_{k=1}^{n} \frac{2k-1}{2n}\varphi(k) \tag{12}$$

, a good measure of Lorenz-inequality in the n-overlap vector $\varphi = (\varphi(1),\ldots,\varphi(n))$. Hence, any increasing function of

$$S = \sum_{k=1}^{n} \frac{2k-1}{2n}\varphi(k) \tag{13}$$

is a good measure of similarity in the n-overlap vector φ. Amongst all these good measures we will now derive the generalized n-overlap Jaccard measure J_n (hence such that $J_2 = J$). For $n = 2$ we see that

$$S = \frac{1}{4}\varphi(1) + \frac{3}{4}\varphi(2)$$

$$= \frac{1}{4}(1-J) + \frac{3}{4}J$$

$$S = \frac{1}{4} + \frac{1}{2}J$$

So

$$\varphi(2) = J = 2\left(S - \frac{1}{4}\right) \tag{14}$$

For $n = 3$ we find

$$S = \frac{1}{6}\varphi(1) + \frac{3}{6}\varphi(2) + \frac{5}{6}\varphi(3)$$

$$= \frac{1}{6}\left(1 - \varphi(2) - \varphi(3)\right) + \frac{3}{6}\varphi(2) + \frac{5}{6}\varphi(3)$$

$$= \frac{1}{6} + \frac{1}{3}\varphi(2) + \frac{2}{3}\varphi(3)$$

So

$$\varphi(2) + 2\varphi(3) = 3\left(S - \frac{1}{6}\right) \tag{15}$$

which extends $J = \varphi(2)$ in case $\varphi(3) = 0$. In general we have

$$S = \frac{1}{2n}\varphi(1) + \frac{3}{2n}\varphi(2) + \frac{5}{2n}\varphi(3) + \ldots + \frac{2n-1}{2n}\varphi(n)$$

$$S = \frac{1}{2n}\left(1 - \varphi(2) - \varphi(3) - \ldots - \varphi(n)\right) + \frac{3}{2n}\varphi(2) + \frac{5}{2n}\varphi(3) + \ldots + \frac{2n-1}{2n}\varphi(n)$$

$$S = \frac{1}{2n} + \frac{1}{n}\varphi(2) + \frac{2}{n}\varphi(3) + \ldots + \frac{n-1}{n}\varphi(n)$$

So

$$\varphi(2)+ 2\varphi(3)+ ...+ (n- 1)\varphi(n)= n\left(S- \frac{1}{2n}\right) \quad (16)$$

extends each previous case $(i = 2,...,n- 1)$ and is an increasing function of S, hence a good measure of similarity in the n-overlap vector φ. We therefore define

$$J_n = \sum_{k=2}^{n} (k- 1)\varphi(k) \quad (17)$$

to be the generalized Jaccard index for the n-overlap (generalizing the classical Jaccard index $J = J_2$ for 2-overlap).

III. Functional expressions for the n-overlap vector φ

III.1 Type and token n-overlap

In this section we will study n-overlap in two versions: type and token n-overlap, defined as follows. Let us have n families $D_1,...,D_n$ consisting of objects that are characterized (indexed) using a certain source set S (e.g. a set of N-grams as types but the representation can be anything). Let us denote, for $k = 1,...,n$

$F(D_k)$ = the set of those sources in S that occur (as representation) in D_k as type (18)

$G(D_k)$ = the family of those occurences of sources (i.e. items) in S that occur (as representation) in D_k as token (19)

Note that, as already remarked in the first section, that $|F(D_k)| \leq |G(D_k)| = |D_k|$ for every $k = 1,...,n$ (of course $|G(D_k)|$ denotes the total number of elements, with their repetition, as they occur in D_k).

Let us now define the n-overlap vectors for type and token.

We denote, for $k = 1,...,n$

$\lambda(k)$= the fraction (with respect to $\left|\bigcup_{i=1}^{n} F(D_i)\right|$) of the sources (from S) that belong to exactly k sets out of the $F(D_1),...,F(D_n)$. (20)

Let us denote by

$$\bigcup_{i=1}^{n} G(D_i)$$

the family consisting of all elements of each $G(D_i)$, with their repetition. Denote, for $k = 1,...,n$

$\xi(k)$= the faction (with respect to $\left|\bigcup_{i=1}^{n} G(D_i)\right|$) of the occurrences of the sources (i.e. the items) that are spread out over exactly k families out of the $G(D_1),...,G(D_n)$. (21)

Let us denote by λ the vector $(\lambda(1),...,\lambda(n))$ and by ξ the vector $(\xi(1),...,\xi(n))$ being two examples of an n-overlap vector φ as defined in the previous sections. It is trivial that we have

$$\sum_{k=1}^{n} \lambda(k) = \sum_{k=1}^{n} \xi(k) = 1 \quad (22)$$

Let us give a simple example where the source set consists of letters of the alphabet.

$$G(D_1) = \{a, b, a, c, d, e\}$$

$$G(D_2) = \{b, d, f, g, b, e, e\}$$

$$G(D_3) = \{b, h, h\}$$

Then

$$F(D_1) = \{a, b, c, d, e\}$$

$$F(D_2) = \{b, d, f, g, e\}$$

$$F(D_3) = \{b, h\}$$

We have

$$\dot{\bigcup}_{i=1}^{3} F(D_i) = \{a, b, c, d, e, f, g, h\}$$

$$\dot{\bigcup}_{i=1}^{3} G(D_i) = \{a, a, b, b, b, b, c, d, d, e, e, e, f, g, h, h\}$$

Hence $\lambda = \left(\frac{5}{8}, \frac{2}{8}, \frac{1}{8}\right)$ and $\xi = \left(\frac{7}{16}, \frac{5}{16}, \frac{4}{16}\right)$. This makes it also obvious that $\lambda \neq \xi$ and hence that it is worthwhile to study both n-overlap vectors. In the sequel we will show relations between these n-overlap vectors λ and ξ.

We will now prove some functional forms for the functions $k \to \lambda(k)$ and $k \to \xi(k)$, $k = 1, \ldots, n$ supposing some functional relations for

(1) the source-item relationship (as they occur in $\dot{\bigcup}_{i=1}^{n} G(D_i)$)

(2) the sizes of the families D_i, i.e. $|D_i|$, the number of elements (with repetition) in the families D_i, $i = 1, \ldots, n$.

At the same time we claim (and we will prove it) that these two attributes are determining n-overlap (both λ and ξ), a fact that has not been remarked before.

III.2 Source-item and families' sizes relationships

In Egghe (2005a) (and many references therein) it is shown that the power law relationship for source-items as well as for sizes of sets (extendable to sizes of families) is the most common and most accepted functional relation. This means the following. We define the overall source-item (i.e. type-token) rank-frequency distribution to be Zipfian:

$$P(r) = \frac{D}{r^{\beta}} \tag{23}$$

where $P(r)$ denotes the fractional density of items in the source on rank-density $r \geq 1$ and where $\beta > 0$.

Next, comparable with city-sizes or journal-sizes we define the families' sizes, also in the rank-frequency form, to be a Zipfian function (not distribution): denote by $|D_k| (k = 1,...,n)$ the number of elements (with repetition) in the family D_k, then, with $\gamma > 0$:

$$h(k) = |D_k| = \frac{C}{k^{\gamma}} \tag{24}$$

(i.e. the families are ranked in decreasing order of $|D_k|$).

III.3 General aspects of n-overlap

In the next subsections we will make distinction between type- and token-n-overlap, but both forms of n-overlap are determined by the same model.

Since (23) expresses the overall probability for a source element to occur (i.e. as an item) and since (24) expresses the number of "possible occurrences" in D_k $(k = 1,...,n)$ we have that an item occurs in exactly k families if and only if, for $k = 1,...,n-1$

$$\frac{D}{r^{\beta}} \frac{C}{k^{\gamma}} \geq 1 > \frac{D}{r^{\beta}} \frac{C}{(k+1)^{\gamma}} \tag{25}$$

, i.e. counting an occurrence as existing from 1 onwards (the right inequality can be deleted for $k = n$, keeping the first one). Relation (25) expresses that the source on rank r appears (or rather "is used") in family D_k (hence also in the families $D_1,...,D_{k-1}$, because of (24)) and not in the family D_{k+1} (hence also not in the families $D_{k+2},...,D_n$, again because of (24)), hence that the source on rank r appears in exactly k sets $F(D_1),...,F(D_k)$ as type and k families $G(D_1),...,G(D_k)$ as token. Relation (25) is equivalent with

$$\frac{(k+1)^{\gamma}}{C} > P(r) = \frac{D}{r^{\beta}} \geq \frac{k^{\gamma}}{C} \tag{26}$$

for $k = 1,...,n-1$ and for $k = n$ we have

$$P(r) = \frac{D}{r^{\beta}} \geq \frac{n^{\gamma}}{C} \tag{27}$$

Before we can proceed further with this general model for n-overlap we first need some general results from source-item informetrics (also called dual informetrics theory).

III.4 General aspects of source-item informetrics

From the general source-item informetrics theory (see Egghe and Rousseau (1990), Egghe (2005a) and references therein) we have that

$$j = g(r) = AP(r) = \frac{AD}{r^{\beta}} \tag{28}$$

is the rank-frequency function of the source-item relationship and that $j \in [1, \rho_m]$ (ρ_m = maximum item density), being the variable of the size-frequency function $f = f(j)$ denoting the density of the sources with item-density j. Here A denotes the total number of items in the system (equalling here the total number of elements $\left| \bigcup_{k=1}^{n} G(D_k) \right|$). We refer again to the above

references for a proof of the following theorem (which is a combination of results of Egghe and of Rousseau).

Theorem III.4.1 :

We have equivalency of

(i)

$$g(r) = \frac{AD}{(1+r')^{\beta}} \tag{29}$$

$\beta > 0$, $r' \in [0,T]$ (hence (28) with $r = 1+r' \geq 1$)

(ii)

$$f(j) = \frac{E}{j^{\alpha}} \tag{30}$$

$\alpha > 1$, $j \in [1,\rho_m]$,

$$\rho_m = \left(\frac{E}{\alpha - 1}\right)^{\frac{1}{\alpha - 1}} = AD \tag{31}$$

and where

$$\alpha = 1 + \frac{1}{\beta}, \tag{32}$$

hence Lotka's power law for $\alpha > 1$. For $\alpha = 1$ (also interesting in this paper as we will see in the sequel) we have the following theorem.

Theorem III.4.2 :

If $\alpha = 1$ we have equivalency of

(i)

$$g(r) = Hq^r \tag{33}$$

$r \in [0, T]$

(ii)

$$f(j) = \frac{E}{j} \tag{34}$$

$j \in [1, \rho_m],\ 0 < q = e^{-\frac{1}{E}} < 1$

i.e. Lotka's power law f with exponent $\alpha = 1$ is equivalent with the exponentially decreasing rank-frequency function g.

The latter theorem can be found in Egghe (2005a) and Egghe and Rousseau (2003). In conclusion, Lotka's power law $f(j) = \frac{E}{j^\alpha}$ with $\alpha \geq 1$ comprises Zipf's power law (28) (if $\alpha > 1$) and an exponentially decreasing rank-frequency function (33) (if $\alpha = 1$). The latter model will also be applied in the connection of n-overlap.

Note :
We could also use that the distribution P (in (23)) and function g (in (28)) is a Mandelbrot function in which case we obtain, in Theorem III.4.1, an equivalency with Lotka's function (30) for even all $\alpha > 0$ (instead of $\alpha > 1$) (cf. Egghe (2005a)). We leave the extension of our model to this case to the reader. The values $0 < \alpha < 1$ are not very common in practise and including them – although allowed – would yield a mixture of different results which would make this paper more confusing to read and understand.

We can now continue the elaboration of the general n-overlap model, expressed in (26) and (27). First we study type n-overlap.

III.5 Type n-overlap

We can now calculate the functional form of the type n-overlap vector $\lambda = (\lambda(1),...,\lambda(n))$.

Theorem III.5.1 :

Under the assumptions (23) and (24) we have that the type n-overlap vector λ has the form (note that $\gamma > 0$ and $\alpha > 1$)

$$\lambda(k) = M\left(\frac{1}{k^{\gamma(\alpha-1)}} - \frac{1}{(k+1)^{\gamma(\alpha-1)}}\right) \tag{35}$$

approximating a power law :

$$\lambda(k) \approx \frac{M\gamma(\alpha-1)}{k^{1+\gamma(\alpha-1)}} \tag{36}$$

if $k = 1,...,n-1$ and

$$\lambda(n) = M\left(\frac{1}{n^{\gamma(\alpha-1)}} - M'\right) \tag{37}$$

where M and M' are constants and where $\alpha > 1$. Hence $\lambda(k)$ decreases in k. This implies that the Lorenz-curve of λ equals

$$L_\lambda\left(\frac{k}{n}\right) = M\left(1 - \frac{1}{(k+1)^{\gamma(\alpha-1)}}\right) \tag{38}$$

, i.e. a constant minus a power law. L_λ increases with γ and with α (i.e. L_λ follows the increases of the Lorenz-curves of h and f).

Proof :

From (26), (27) and (28) we have

$$\frac{A(k+1)^{\gamma}}{C} > j \geq \frac{Ak^{\gamma}}{C} \tag{39}$$

for $k = 1,...,n-1$ and (since $j \leq \rho_m$ always)

$$\rho_m \geq j \geq \frac{An^{\gamma}}{C} \tag{40}$$

for $k = n$.

The size-frequency function f is needed in the distinction between type- and token-n-overlap. For type-n-overlap we have, in general notation, that

$$\frac{1}{T}\int_a^b f(j)dj \tag{41}$$

is the fraction of sources (T = total number of sources in the system, equalling here $\left|\bigcup_{k=1}^{n} F(D_k)\right|$) with item density j between a and b. For the values given in (39) and (40) we hence have that (41) gives the fraction of the sources (i.e. types) that belong to exactly k sets $F(D_1),...,F(D_k)$, hence $\lambda(k)$, $k = 1,...,n$. More concretely we have for $k = 1,...,n-1$:

$$\lambda(k) = \frac{1}{T}\int_{\frac{Ak^{\gamma}}{C}}^{\frac{A(k+1)^{\gamma}}{C}} \frac{E}{j^{\alpha}}dj \tag{42}$$

and

$$\lambda(n) = \frac{1}{T}\int_{\frac{An^{\gamma}}{C}}^{\rho_m} \frac{E}{j^{\alpha}}dj \tag{43}$$

where $\alpha > 1$, using Theorem III.4.1. Hence, for $k = 1,...,n-1$ (since $\alpha > 1$)

$$\lambda(k) = \frac{EC^{\alpha-1}}{T(\alpha-1)A^{\alpha-1}}\left(\frac{1}{k^{\gamma(\alpha-1)}} - \frac{1}{(k+1)^{\gamma(\alpha-1)}}\right)$$

which is (35) for the appropriate choice of M (its exact form is of no importance here since λ is normalized). Formula (37) is found similarly.

Hence, using that

$$\frac{1}{k^{\gamma(\alpha-1)}} - \frac{1}{(k+1)^{\gamma(\alpha-1)}} \gg -\left(\frac{1}{k^{\gamma(\alpha-1)}}\right)',$$

the derivative of the function $k \to \frac{1}{k^{\gamma(\alpha-1)}}$, we find (36), a decreasing power function (since γ and $\alpha - 1$ are > 0).

By definition of the Lorenz-curve of λ (formula (11) for $\varphi = \lambda$) we have, for $k = 1,...,n-1$ that (since λ decreases in k, since $\alpha > 1$)

$$L_\lambda\left(\frac{k}{n}\right) = M\left(1 - \frac{1}{(k+1)^{\gamma(\alpha-1)}}\right)$$

(and, as always, $L_\lambda(0) = 0$ and $L_\lambda(1) = 1$).

From this it follows that L_λ is an increasing function of γ and of α, a fact that is already proved in Egghe (2005c), see also Egghe (2005a): a Lorenz-curve of a decreasing power law is increasing in the exponent of the power law. From this it also follows readily that L_λ is increasing with L_f and with L_h, the Lorenz-curves of the Lotka-function f (expressing the source-item size-frequency relationship) and of the Zipf-function h (expressing the family rank-size relationship). This finishes the proof of the functional form (and its properties) of the type-n-overlap vector λ. ~

We now turn our attention to token-n-overlap.

III.6 Token-n-overlap

Now we can calculate the functional form of the n-overlap vector $\xi = (\xi(1),...,\xi(n))$.

Theorem III.6.1 :

Under the assumptions (23) and (24) we have that the token-n-overlap vector ξ has the form, for $\alpha \neq 2$

$$\xi(k) = N\left(\frac{1}{k^{\gamma(\alpha-2)}} - \frac{1}{(k+1)^{\gamma(\alpha-2)}}\right) \tag{44}$$

approximating a power law

$$\xi(k) \approx \frac{N\gamma(\alpha-2)}{k^{1+\gamma(\alpha-2)}} \tag{45}$$

for $k = 1,...,n-1$ and

$$\xi(n) = N\left(\frac{1}{n^{\gamma(\alpha-2)}} - N'\right) \tag{46}$$

where N and N' are constants. Hence $\xi(k)$ is decreasing in k if $\alpha > 2$ and increasing if $\alpha < 2$. The case $\alpha = 2$ is a continuous limiting case of the cases $\alpha < 2$ and $\alpha > 2$ covered in (44) and hence is not treated here, for the case of simplicity. The Lorenz-curve of ξ equals (for $\alpha > 2$, the most common case – see Egghe (2005a), Chapter I)

$$L_{\xi}\left(\frac{k}{n}\right) = N\left(1 - \frac{1}{(k+1)^{\gamma(\alpha-2)}}\right) \tag{47}$$

, i.e. a constant minus a power law. Again L_ξ increases with γ and α, i.e. L_ξ follows the increases of the Lorenz-curves of h and f.

Furthermore, in this case

$$L_\gamma > L_\xi \tag{48}$$

everywhere (except of course in $(0,0)$ and $(1,1)$). This means that the type-n-overlap vector is more concentrated (unequal) than the token-n-overlap vector.

Proof :

For token-n-overlap we have, in general notation, that

$$\frac{1}{A}\int_a^b jf(j)dj \tag{49}$$

is the fraction of items (A = total number of items in the system, equalling here $\left|\sum_{k=1}^{n} G(D_k)\right| = \left|\sum_{k=1}^{n} D_k\right|$) with item density between a and b. For the values given in (39) and (40) we hence have that (49) gives the fraction of the items (i.e. tokens) that belong to (or are spread out over) exactly k families $G(D_1),...,G(D_k)$, hence $\xi(k)$, $k = 1,...,n$. More concretely we have for $k = 1,...,n-1$:

$$\xi(k) = \frac{1}{A}\int_{\frac{Ak^\gamma}{C}}^{\frac{A(k+1)^\gamma}{C}} \frac{E}{j^{\alpha-1}}dj \tag{50}$$

and

$$\xi(n) = \frac{1}{A}\int_{\frac{An^\gamma}{C}}^{\rho_m} \frac{E}{j^{\alpha-1}}dj \tag{51}$$

where $\alpha > 1$, using Theorem III.4.1 and where we suppose (as indicated above) $\alpha \neq 2$. Hence, for $k = 1,...,n-1$ we have

$$\xi(k) = \frac{EC^{\alpha-2}}{(\alpha-2)A^{\alpha-1}}\left(\frac{1}{k^{\gamma(\alpha-2)}} - \frac{1}{(k+1)^{\gamma(\alpha-2)}}\right)$$

which is (44) for the appropriate choice of N (again its exact form is of no importance since ξ is normalized). Formula (46) is found similarly.

Hence, using that

$$\frac{1}{k^{\gamma(\alpha-2)}} - \frac{1}{(k+1)^{\gamma(\alpha-2)}} \approx -\left(\frac{1}{k^{\gamma(\alpha-2)}}\right)',$$

the derivative of the function $k \to \frac{1}{k^{\gamma(\alpha-2)}}$, we find (45). The properties of L_ξ, the Lorenz-curve of ξ (a decreasing vector for $\alpha > 2$) are proved in the same way as in the previous theorem (note that again L_ξ increases in γ and α).

As to the comparison of L_λ and L_ξ we have the following result: by (38), L_λ is the normalized function of

$$\frac{k}{n} \to 1 - \frac{1}{(k+1)^{\gamma(\alpha-1)}}$$

while L_ξ, by (47), is the normalized function of

$$\frac{k}{n} \to 1 - \frac{1}{(k+1)^{\gamma(\alpha-2)}}$$

we have that, since $\gamma(\alpha-1) > \gamma(\alpha-2) > 0$ that

$$L_\lambda > L_\xi$$

everywhere (except in $0 = (0,0)$ and $(1,1)$. This also follows from the approximate results (36) and (45) and the mentioned theorems on Lorenz-curves of power laws in the proof of the previous theorem (see Egghe (2005a) or Egghe (2005c)). This finishes the proof of the results on the token-n-overlap vector ξ. ~

Remark 1 :

Also in subsection III.1 an example was given where $L_\lambda > L_\xi$: the example $\lambda = \left(\frac{5}{8}, \frac{2}{8}, \frac{1}{8}\right)$, $\xi = \left(\frac{7}{16}, \frac{5}{16}, \frac{4}{16}\right)$. Since $\frac{5}{8} > \frac{7}{16}$ and $\frac{7}{8} > \frac{12}{16}$ we indeed have $L_\lambda > L_\xi$. Because of the above theorem, this result is generally valid in the case of power functions for P (hence g) and h (for $\alpha > 2$). There are, however, other cases where this result is not true.

Example :

$$G(D_1) = \{a, a, a, b\}$$
$$G(D_2) = \{b\}$$

Hence

$$F(D_1) = \{a, b\}$$
$$F(D_2) = \{b\}$$

and

$$F(D_1) \dot{\cup} F(D_2) = \{a, b\}$$
$$G(D_1) \dot{\cup} G(D_2) = \{a, a, a, b, b\}$$

Hence $\xi(1) = \frac{3}{5} > \frac{1}{2} = \lambda(1)$, contradicting $L_\lambda > L_\xi$. In fact, since $n = 2$ we have here $L_\xi > L_\lambda$!

Remark 2 :

If, in the above Theorem III.6.1, $\alpha \leq 2$ then we still can use formula (47) for L_ξ but now L_ξ is constructed for an increasing vector ξ (as in (44)). Hence L_ξ is now a convexly increasing function below the first bisectrix $y = x$. Alternatively, we can use its "concave equivalent" by reordering ξ as the vector $(\xi(n),...,\xi(1))$ which is now decreasing (and positive since $N < 0$ now). We leave these constructions to the reader.

The same remark could be made in Theorem III.5.1 if $0 < \alpha < 1$ where, in this case, also the vector λ is increasing (note that now $M < 0$). The same remark as above for ξ can now be made for λ which execution is also left to the reader.

III.7 Type- and token-n-overlap in case $\alpha = 1$

In this case we assume, following Theorem III.4.2 that $P(r)$ and $g(r)$ are exponentially decreasing functions. We have the following results (only $k = 1,...,n-1$ is given, the most important cases since the vectors are normalized). By (41), (39) and (34) we have

$$\lambda(k) = \frac{1}{T}\int_{\frac{Ak^\gamma}{C}}^{\frac{A(k+1)^\gamma}{C}} \frac{E}{j}\,dj$$

$$\lambda(k) = \frac{E}{T}\gamma \ln\left(\frac{k+1}{k}\right) \tag{52}$$

a decreasing function of k. The Lorenz-curve of λ hence has the form $(k = 1,...,n)$

$$L_\lambda\left(\frac{k}{n}\right) = \frac{\ln(k+1)}{\ln(n+1)} \tag{53}$$

By (49), (39) and (34) we have

$$\xi(k) = \frac{1}{A}\int_{\frac{Ak^{\gamma}}{C}}^{\frac{A(k+1)^{\gamma}}{C}} E dj$$

$$\xi(k) = \frac{E}{C}\left[(k+1)^{\gamma} - k^{\gamma}\right]$$

, now increasing in k! Its Lorenz-curve (now starting with $\xi(n),\ldots$) is

$$L_{\xi}\left(\frac{k}{n}\right) = \frac{(n+1)^{\gamma} - (n-k+1)^{\gamma}}{(n+1)^{\gamma} - 1} \qquad (54)$$

We close this section by replacing the families' rank-size function h by a decreasing exponential one. We will explain in the sequel why we also consider this case.

III.8 Type- and token-n-overlap in case h is an exponential function

If we replace h in (24) by a decreasing exponential function of the form

$$h(k) = |D_k| = Cb^k, \qquad (55)$$

where $0 < b < 1$, we have that condition (25) for a source on rank r to be used in exactly k families $D_1,\ldots,D_k$ (i.e. to appear in $F(D_1),\ldots,F(D_k)$ as type and to be spread out over the k families $G(D_1),\ldots,G(D_k)$ as token) now reads:

$$P(r)Cb^k \geq 1 > P(r)Cb^{k+1}$$

So

$$\frac{1}{C}b^{-(k+1)} > P(r) \geq \frac{1}{C}b^{-k}$$

Hence the same argument as in (28) now leads to the following variants of (39) and (40):

$$\frac{A}{C}b^{-(k+1)} > j \geq \frac{A}{C}b^{-k} \tag{56}$$

For the type-n-overlap vector λ we now have (same argument as in Theorem III.5.1), for $k = 1,\ldots,n-1$

$$\lambda(k) = \frac{1}{T}\int_{\frac{A}{C}b^{-k}}^{\frac{A}{C}b^{-(k+1)}} \frac{E}{j^{\alpha}}\,dj$$

(since λ is normalized we do not deal with $\lambda(n)$). Hence

$$\lambda(k) = M\left[b^{k(\alpha-1)} - b^{(k+1)(\alpha-1)}\right] \tag{57}$$

where M is a normalizing factor. λ is decreasing in k (for $\alpha > 1$).

For the token-n-overlap vector ξ we have (same argument as in Theorem III.6.1), for $k = 1,\ldots,n-1$

$$\xi(k) = \frac{1}{A}\int_{\frac{A}{C}b^{-k}}^{\frac{A}{C}b^{-(k+1)}} \frac{E}{j^{\alpha-1}}\,dj$$

$$\xi(k) = N\left[b^{k(\alpha-2)} - b^{(k+1)(\alpha-2)}\right] \tag{58}$$

where N is a normalizing factor. ξ is decreasing in k for $\alpha > 2$ (the most common case as already remarked above). As in Theorems III.5.1 and III.6.1, we can approximate (57) and (58) by

$$\lambda(k) \approx -M(\ln b)(\alpha-1)b^{k(\alpha-1)} \tag{59}$$

(where $-M(\ln b)(\alpha-1) > 0$ for $\alpha > 1$) and

$$\xi(k) \approx -N(\ln b)(\alpha-2)b^{k(\alpha-2)} \qquad (60)$$

(where $-N(\ln b)(\alpha-2) > 0$ for $\alpha > 2$), hence decreasing exponential models.

The results in subsections III.5, III.6 and III.8 yield approximate power models and decreasing exponential models for the n-overlap vectors λ and ξ. Both models were also studied on the data in Table 1 of Hood and Wilson (2003), with a (statistical) preference for the exponential model.

IV. The influence of hierarchy on the n-overlap vector

Objects in the families $D_1,...,D_n$ can be represented (indexed) using different source sets S (as indicated above). It is even so that in many cases one can choose the source set amongst a large variety of indexing refinements. A well-known example are the N-grams (see e.g. Egghe (2000) or Egghe (2005b)). N-grams are concatenations of N letters (more generally N symbols) determined by author and/or title words. It is clear that, when $N \to \infty$ increases, the indexing becomes finer and less documents are represented by a particular N-gram.

Other hierarchical indexing systems are common, say the linking of a document to a decimal number, representing a topic, and the longer the decimal number, the finer the indexing (comparable with N-grams). Examples are the UDC (Universal Decimal Classification), the Dewey system and many other ones.

To fix the ideas we will always use the terminology of N-grams, partially also because we will refer to some papers we have written on this topic (e.g. the ones mentioned above) and we will use results that were proved in these papers. But the reader should bear in mind that

all results on n-overlap of N-grams (particularly on variable N) can be extended to all other hierarchical systems (e.g. yielding results on variable hierarchical level).

We have two subsections in this Section IV. The first one deals with the application of results from Section III to N-grams (hence to general hierarchical systems). The second subsection deals with the influence of the indexing level (i.e. of N in an N-gram) on the Lorenz order determined by the n-overlap vector.

IV.1 Applications of the results of Section III to N-grams

It is clear that the general results of Section III can be applied to the case that the source set S consists of N-grams. Basic in this application are the assumptions (23) and (24) (on the source-item relationship and the families' sizes, respectively). It is clear that (24) can still be used since this has nothing to do with the used indexing method. For (23), we remark that our arguments (essentially formulae (39) for type- and token-n-overlap, (41) for type-n-overlap and (49) for token-n-overlap) only use the size-frequency function f.

We are lucky with this remark since, as proved in Egghe (2005a,b), the rank-frequency distribution P (called here P_N to indicate that we deal with N-grams) is extremally complicated while its size-frequency equivalent (denoted f_N) is relatively simple. In Egghe (2004a,b) we prove, based on (23) for 1-grams (i.e. the rank-frequency distribution for single symbols is supposed to be Zipfian), that the size-frequency distribution of N-grams satisfies

$$f_N(j) = \frac{E}{j^{\alpha}} \ln^{N-1}\left(\frac{\rho_m}{j}\right) \qquad (61)$$

for $j \in [1, \rho_m]$ (ρ_m is again the maximal N-gram item-density) and

$$\alpha = 1 + \frac{1}{\beta} \qquad (62)$$

(as in the pure Zipfian case – see (32)).

Although relatively simple, formula (61) is not exactly a power law, a fact that we used in the previous section (see Theorem III.4.1, formula (30)).

But as is easily seen, function (61) has the shape of a power law, i.e. convexly decreasing and where we can put

$$f_N(j) \approx \frac{E_N}{j^{\alpha_N}} \tag{63}$$

where E_N is a constant (dependent on N) and where α_N is an exponent which is increasing in N. This follows from the fact that, in (61), f_N depends (besides on $\frac{1}{j^{\alpha}}$) also on $\left(\ln \frac{1}{j}\right)^{N-1}$ which results in the fact that f_N decreases more rapidly for increasing N.

In Theorems III.5.1 (for the type-n-overlap vector λ) and III.6.1 (for the token-n-overlap vector ξ) we showed, now using (63), that (now denoting λ_N respectively ξ_N to show the N-dependence)

$$\lambda_N(k) \approx \frac{M_N \gamma(\alpha_N - 1)}{k^{1+\gamma(\alpha_N - 1)}} \tag{64}$$

$$\xi_N(k) \approx \frac{M'_N \gamma(\alpha_N - 2)}{k^{1+\gamma(\alpha_N - 2)}} \tag{65}$$

and from the conclusions of both theorems we have that the corresponding Lorenz-curves L_{λ_N} and L_{ξ_N} increase with increasing α_N, hence (since α_N is a strictly increasing function of N) we have that L_{λ_N} and L_{ξ_N} are strictly increasing in N.

This approximate result, based on power laws, will now be reproved exactly and in a general way, only using hierarchical aspects (of, in this case, N-grams) but for the token-n-overlap vector ξ_N only.

IV.2 Exact proof of the influence of hierarchy on the token-n-overlap vector ξ_N

Independent of the underlying rank- or size-frequency laws for source-items or for families' sizes we have the following results.

Theorem IV.2.1 :

For all $k = 1,...,n$ and $N \in \mathbb{N}$, we have

$$\sum_{j=1}^{k} \xi_{N-1}(j) \geq \sum_{j=1}^{k} \xi_N(j) \tag{66}$$

and

$$\sum_{j=k}^{n} \xi_{N-1}(j) \leq \sum_{j=k}^{n} \xi_N(j) \tag{67}$$

Proof :

It is clear that (67) follows from (66) since

$$\sum_{j=1}^{n} \xi_{N-1}(j) = \sum_{j=1}^{n} \xi_N(j) = 1 \tag{68}$$

for all $N \in \mathbb{N}$.

To prove (66), note that

$$\xi_N(j) = \frac{\text{total number of } N\text{-gram tokens that are spread out over exactly } j \text{ families } G_N(D_1),...,G_N(D_j)}{\left| \bigcup_{k=1}^{n} G_N(D_k) \right|} \tag{69}$$

(again we replaced G in (19) by G_N to show the N-dependence).

Now, for every N-gram token that belongs to j families $G_N(D_1),...,G_N(D_j)$ it is clear that its $(N-1)$- gram equivalent (i.e. the truncation of this N-gram to the corresponding $(N-1)$- gram (e.g. ABCD $(N=4)$® ABC$(N=3)$) belongs to $j^{\prime}$ families $G_{N-1}(D_1),...,G_{N-1}(D_{j^{\prime}})$, where $j^{\prime} \geq j$ ($j^{\prime}$ can be > j e.g. in the above example ABCE ® ABC).

Let now $j \in \{1,...,k\}$, $k=1,...,n$ fixed. If $j^{\prime} \in \{1,...,k\}$ then this $(N-1)$- gram is counted in the numerator of one of the $j^{\prime}$th term in $\sum_{j=1}^{k} \xi_{N-1}(j)$ in the same way as this N-gram is counted in the numerator of the jth term in $\sum_{j=1}^{k} \xi_N(j)$.

Since $j^{\prime} \geq j$ $j \in \{1,...,k\}$ it can also be that $j^{\prime} > k$. In that case this $(N-1)$- gram is not counted in any nominator of a term in $\sum_{j=1}^{k} \xi_{N-1}(j)$ while the N-gram still is counted in the nominator of the jth term in $\sum_{j=1}^{k} \xi_N(j)$. Now, if $j > k$ then also $j^{\prime} \geq j > k$ and both N-gram and $(N-1)$- gram are not counted in $\sum_{j=1}^{k} \xi_{N-1}(j)$ nor in $\sum_{j=1}^{k} \xi_N(j)$. Hence, in general, the nominator in $\sum_{j=1}^{k} \xi_{N-1}(j)$ is smaller than the nominator in $\sum_{j=1}^{k} \xi_N(j)$. But their denominators are the same since we deal here with token-n-overlap:

$$\left| \bigcup_{k=1}^{n} G_N(D_k) \right| = \left| \bigcup_{k=1}^{n} G_{N-1}(D_k) \right| = \left| \bigcup_{k=1}^{n} D_k \right| \qquad (70)$$

Hence (66) and hence also (67) is proved. ~

<u>Corollary IV.2.1</u> :

(i) Let the token n-overlap vectors $\xi_N = (\xi_N(1),...,\xi_N(n))$ and $\xi_{N-1} = (\xi_{N-1}(1),...,\xi_{N-1}(n))$ be decreasing. Then we have that

$$L_{\xi_{N-1}} \pounds L_{\xi_N} \tag{71}$$

everywhere ($L_{\ldots}$ denotes the Lorenz-curve of the vector mentioned in the index).

(ii) Let the token n-overlap vectors ξ_{N-1} and ξ_N be increasing. Then we have that

$$L_{\xi_{N-1}} \,^3\, L_{\xi_N} \tag{72}$$

everywhere.

Proof :

(i) For decreasing vectors the construction of the Lorenz-curve (explained in Section II) is executed from the first coordinate onwards (see (11) with $\varphi = \xi_{N-1}$ or $\varphi = \xi_N$). It is clear that (66) now implies (71).

(ii) For increasing vectors the Lorenz-curves are constructed from the nth coordinate onwards (since $(\xi_N(n),...,\xi_N(1))$ is decreasing and the same for $(\xi_{N-1}(n),...,\xi_{N-1}(1))$. Now (67) implies (72). ~

Theorem IV.2.1 is false for the type-n-overlap vectors λ_N , λ_{N-1}. Indeed, formula (66), for $k = 1$:

$$\xi_{N-1}(1) \pounds \xi_N(1) \tag{73}$$

is false for λ_N , λ_{N-1} as the next examples shows : $N = 2$, $n = 2$

$$F_2(D_1) = \{ab, ac, ad, bc\}$$

$$F_2(D_2) = \{ab, ac, ad\}$$

so

$$F_1(D_1) = \{a, b\}$$
$$F_1(D_2) = \{a\}$$

and

$$F_2(D_1) \dot{E} F_2(D_2) = \{ab, ac, ad, bc\}$$
$$F_1(D_1) \dot{E} F_1(D_2) = \{a, b\}$$

Hence $\lambda_2(1) = \frac{1}{4} < \lambda_1(1) = \frac{1}{2}$, contradicting (73). Note that λ_1 and λ_2 are increasing.

One could wonder if, for λ_N, λ_{N-1}, the opposite result could be true: this would then require that, always

$$\lambda_N(1) \pounds \lambda_{N-1}(1) \qquad (74)$$

(as is the case in the previous example). This is not true either as the next example shows.

$$F_2(D_1) = \{ab, ac\}$$
$$F_2(D_2) = \{ad, ae, bc\}$$

so

$$F_1(D_1) = \{a\}$$
$$F_1(D_2) = \{a, b\}$$

and

$$F_2(D_1) \cap F_2(D_2) = \{ab, ac, ad, ae, bc\}$$

$$F_1(D_1) \cap F_1(D_2) = \{a, b\}$$

Hence $\lambda_2(1) = 1 > \lambda_1(1) = \frac{1}{2}$.

Decreasing n-overlap vectors are the most natural ones in practise (we refer to the arguments given in Section I and to Table 1). In view of the result in Corollary IV.2.1 we hence have that the Lorenz-concentration of the token-n-overlap vector ξ_N is higher than the one of the token-n-overlap vector ξ_{N-1}. Hence L_{ξ_N} increases with N and hence the vector $(\xi_N(1), ..., \xi_N(n))$ is more unequal than the vector $(\xi_{N-1}(1), ..., \xi_{N-1}(n))$.

Of course this is not an absolute result since also increasing n-overlap vectors exist. One even has n-overlap vectors that are neither increasing nor decreasing.

Examples :

1. Database

1	aa	ba	ca
2	ab	bb	cb
3	ac	bc	cc

$\xi_1 = (0, 0, 1)$, $\xi_2 = (1, 0, 0)$

2. Database

1	aa	ab	ac
2	ba	bb	bc
3	ca	cb	cc
4	ad	ae	af
5	bd	be	bf
6	cd	ce	cf

$$\xi_1 = (0,1,0,0,0,0), \quad \xi_2 = (1,0,0,0,0,0)$$

Note, however, that Theorem IV.2.1 is valid for any token-n-overlap vector ξ_N, ξ_{N-1}, whether they are increasing, decreasing or not. Note also that the argument developed in the proof of this theorem can be repeated for any hierarchical system and where then ξ_{N-1} is replaced by the token-n-overlap vector of the rougher system and ξ_N is replaced by the token-n-overlap vector of the finer system.

V. Conclusions

In this paper we studied n-overlap, i.e. given n databases (sets, libraries,…), we study the vector $\varphi = (\varphi(1),...,\varphi(n))$, where $\varphi(i)(i = 1,...,n)$ denotes the fraction of objects (documents) that belong to exactly i databases. Objects are represented through a set of sources (i.e. indexing descriptions such as N-grams, decimal classifications or even ISBNs). We study the vector φ as a type-n-overlap vector (then denoted as λ) and as a token-n-overlap vector (then denoted as ξ).

We presented a generalization of the classical Jaccard measure J (measuring 2-overlap) to the case of n-overlap :

$$J_n = \sum_{k=2}^{n} (k-1)\varphi(k) \qquad (75)$$

where φ is the vector described above and we showed that $J_2 = J$. J_n is a measure of n-overlap similarity.

Next we present exact definitions of the type and token n-overlap vectors λ and ξ and give functional forms (power laws and exponential laws) for their coordinates. These are based on generally accepted assumptions on the source-item relationship (supposed to be Zipfian) and

on the size distribution of the n databases (Zipfian or decreasing exponential). We prove that both vectors λ and ξ yield Lorenz-curves L_λ and L_ξ which increase if the Lorenz-curves of the source-item distribution and/or the one of the size distribution increases. We also show that, in most cases, the vectors λ and ξ are decreasing, confirming experimental data. We further show that

$$L_\lambda > L_\xi \tag{76}$$

in these cases, i.e. the type-n-overlap vector λ is more concentrated than the token-n-overlap vector ξ. We also present examples where (76) is not valid.

Finally the influence of the indexing hierarchy (as e.g. expressed by $N\hat{I}\,¥$ in N-gram indexing) on the n-overlap vectors λ and ξ is discussed. In general we show (using some approximations) that the λ-vector (now denoted λ_N) and that the ξ-vector (now denoted ξ_N) become more concentrated (in the sense of Lorenz) if N increases. These approximate results are proved in an exact way (without any approximations) for the decreasing token-n-overlap vector ξ_N.

References

L. Egghe (2000). The distribution of N-grams. Scientometrics 47(2), 237-252, 2000.

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

L. Egghe (2005b). The exact rank-frequency function and size-frequency function of N-grams and N-word phrases with applications. Mathematical and Computer Modelling 41, 807-823, 2005.

L. Egghe (2005c). Zipfian and Lotkaian continuous concentration theory. Journal of the American Society for Information Science and Technology 56(9), 935-945, 2005.

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

L. Egghe and R. Rousseau (1990b). Elements of concentration theory. IN: Informetrics 89/90. Proceedings of the second international Conference on Bibliometrics, Scientometrics and Informetrics, London (Canada), L. Egghe and R. Rousseau (eds.). Elsevier, Amsterdam, 97-137, 1990.

L. Egghe and R. Rousseau (2001). Elementary Statistics for effective Library and Information Management. ASLIB, London (UK), 2001 (available from Europa Publications).

L. Egghe and R. Rousseau (2003). Size-frequency and rank-frequency relations, power laws and exponentials: a unified approach. Progress in Natural Science 13(6), 478-480, 2003.

L. Egghe and R. Rousseau (2005). 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. Information Processing and Management, to appear, 2005.

M. Gluck (1990). A review of journal coverage overlap with an extension to the definition of overlap. Journal of the American Society for Information Science 41(1), 43-60, 1990.

W.W. Hood and C.S. Wilson (2003). Overlap in bibliographic databases. Journal of the American Society for Information Science and Technology 54(12), 1091-1103, 2003.

G. Salton and M.J. McGill (1983). Introduction to modern Information Retrieval. McGraw-Hill, New York, 1983.