Strong similarity measures for ordered sets of documents in information retrieval

EGGHE, Leo & Michel, Chr. (2002) Strong similarity measures for ordered sets of documents in information retrieval. In: Information Processing & Management, 38(6). p. 823-848.

DOI: 10.1016/S0306-4573(01)00051-6
Handle: http://hdl.handle.net/1942/781

STRONG SIMILARITY MEASURES FOR ORDERED SETS OF DOCUMENTS IN INFORMATION RETRIEVAL

by

L. Egghe — LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
UIA, Universiteitsplein 1 , B-2610 Antwerpen (Wilrijk), Belgium
leo.egghe@luc.ac.be

C. Michel — CEM-GRESIC, MSHA, D.U. Bordeaux III, Esplanade des Antilles, F-33607, PESSAC Cedex, France
Christine.Michel@montaigne.u-bordeaux.fr

ABSTRACT

A general method is presented to construct ordered similarity measures (OS-measures), i.e. similarity measures for ordered sets of documents (as e.g. being the result of an IR-process), based on classical, well-known similarity measures for ordinary sets (measures such as Jaccard, Dice, Cosine or overlap measures). To this extent, we first present a review of these measures and their relationships.

The method given here to construct OS-measures extends the one given by Michel in a previous paper so that it becomes applicable on any pair of ordered sets. Concrete expressions of this method, applied to the classical similarity measures, are given.

[1] Permanent address.

Some of these measures are then tested in the IR-system Profil-Doc. The engine SPIRIT© extracts ranked document sets in 3 different contexts, each for 550 requests. The practical useability of the OS-measures is then discussed based on these experiments.

I. Introduction.

Similarity measures for sets of objects (such as documents) are well-known and well-used in the IR literature. They have become standard tools, that are featuring in any good monograph on IR. Relatively recent monographs dealing with these measures are Boyce, Meadow and Kraft (1995), Tague-Sutcliffe (1995), Grossman and Frieder (1998) and Losee (1998). Of course the "standard" books on IR, Salton and Mc Gill (1987) and van Rijsbergen (1979) should also be mentioned here.

In general one has a "universe" Ω from which subsets A,B, ... are generated. In IR, Ω usually is a database of documents and the subsets A,B, ... are the result of an IR-process that started after presenting a query to the system. For any $A,B \subset \Omega$, a similarity measure D gives a positive number D(A,B) that expresses the "degree" of similarity between the two sets. Further on we will go into the properties that good similarity measures should have but we can already understand that D(A,B) should reach its minimal value (usually 0) if $A \cap B = \varnothing$ and should reach its maximal value if $A=B \neq \varnothing$. This maximal value usually is 1 if D is normalised. If this is the case then 1-D is a dissimilarity measure. In this paper we will restrict ourselves to the study of similarity measures.

Examples of important similarity measures are ($A,B \subset \Omega$, $A,B \neq \varnothing$, $|A|$ denotes the cardinality of A).

Jaccard's index J

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

Recall R and Precision P

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

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

Note that here the roles of A and B cannot be interchanged. In IR, A=ret, the set of retrieved documents and B=rel, the set of relevant documents. For this reason, R and P are non-symmetric ($R(A,B) \neq R(B,A)$ and $P(A,B) \neq P(B,A)$). Note that J is symmetric. In most IR cases, R and P are related : the R-P curves (cf. Van Rijsbergen (1979), Salton and Mc Gill (1987)) are decreasing. An ideal search, however should have high values of R and P. There is a similarity measure that expresses this, namely the harmonic average of R and P :

$$E = \frac{2}{\frac{1}{R}+\frac{1}{P}} . \tag{4}$$

It is readily seen that

$$E(A,B) = \frac{2|A \cap B|}{|A|+|B|} , \tag{5}$$

which is also known as Dice's index. Note that E is symmetric. (4) and (5) can be generalized.

(Generalized) Dice Index

For $\alpha \in]0,1[$, we define

$$E_\alpha = \frac{1}{\frac{\alpha}{P}+\frac{1-\alpha}{R}} \tag{6}$$

which reduces to (4) for $\alpha = \frac{1}{2}$. Now

$$E_\alpha(A,B) = \frac{|A\cap B|}{\alpha|A|+(1-\alpha)|B|} \tag{7}$$

Formula (6) appears in Salton and McGill (1987), Van Rijsbergen (1979) and Tague-Sutcliffe (1995) but in the form $1-E_\alpha$. It is not logical to mix similarity and dissimilarity measures ; so we use E_α instead of $1-E_\alpha$. In Tague, one also finds the (apparently) to (4) related

$$\frac{1}{\frac{1}{R}+\frac{1}{P}-1} \tag{8}$$

However, as is readily seen from (2) and (3), (8) is nothing else than J, the Jaccard index.

Cosine measure.

If we take the geometric average of R and P, we obtain another symmetric similarity measure:

$$Cos = \sqrt{RP} = \frac{|A\cap B|}{\sqrt{|A||B|}} \tag{9}$$

The name Cos comes from cosine and will be explained later, when applying (9) for vectors.

In Boyce, Meadow and Kraft (1995) one can find another measure that can be formulated (in our notation as follows) :

$$\frac{|A\cap B|}{\sqrt{|A|^2+|B|^2}} \tag{10}$$

They call it the cosine coefficient. In view of the above (9) and the explanation with vectors (to follow) we cannot accept this name for (10). In addition, all measures encountered so far are normalized (i.e. their maximal value, obtained if $A=B\neq\varnothing$, is 1). This is not the case in (10). The normalized form of (10) is

$$N(A,B) = \sqrt{2}\,\frac{|A\cap B|}{\sqrt{|A|^2+|B|^2}} \tag{11}$$

We will investigate this measure further on (it will be shown that N has good qualities). Note that N is symmetric.

Overlap measures O_1 and O_2

In Boyce, Meadow and Kraft (1995) and Van Rijsbergen (1979) one finds the following overlap measure

$$O_1(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} \tag{12}$$

We did not find a reference for the next measure :

$$O_2(A,B) = \frac{|A \cap B|}{\max(|A|,|B|)} \tag{13}$$

but O_2 will turn out to have better properties than O_1. Furthermore, max is as good as min to be used ; therefore O_2 is worth to be studied.

It follows from elementary calculations that

$$J \leq O_2 \leq E \leq N \leq Cos \leq O_1 \tag{14}$$

and

$$O_2 \leq R, P, E_\alpha \leq O_1 \tag{15}$$

In the next section we will study general properties of these measures, i.e. properties that good similarity measures should have. This will lead to the definition of weak and strong similarity measures.

The third section presents a general method of constructing similarity measures for ordered sets (ordered similarity measures or OS-measures) from similarity measures for ordinary sets as defined above. It turns out that the method can be applied to most of the measures defined

above : J, E (in fact all E_α), Cos, N and O_2. The problem of constructing OS-measures for the other similarity measures (R, P, O_1) is left as an open problem ; the authors intend to study this in the near future. The obtained general method is then applied to concrete strong similarity measures and concrete formulae are given.

The last section is then devoted to experiments. Using the IR-system Profil-Doc and the engine SPIRIT©, we obtain ranked document sets for 550 requests in 3 different contexts, which are then compared (using OS-measures) with a standard neutral ordered set. Graphs are constructed and qualitative conclusions on the used OS-measures are given.

General note on the interpretation of these similarity measures for the comparison of vectors.

All of the above measures compare couples of sets. As such this is very interesting and applicable in IR. However, there is an important part of IR that considers documents and queries as vectors ; each coordinate (which we will restrict here to have values 1 or 0) expresses the fact that a certain key word does or does not belong to the document or the query. Let us limit our attention to documents. In this setting a document is then "replaced" by its vector $\vec{d} = (d_1,...,d_n)$, where each $d_i \in \{0,1\}$, $i=1,...,n$. The above measures can be used to measure the similarity between two such vectors $\vec{d} = (d_i)_{i=1,...,n}$ and $\vec{d'} = (d'_i)_{i=1,...,n}$. This goes as follows : define

$$D = \{i \in \{1,...,n\} \| d_i = 1\} \tag{16}$$

$$D' = \{i \in \{1,...,n\} \| d'_i = 1\} \tag{17}$$

Then

$$|D \cap D'| = <\vec{d}, \vec{d'}>$$

$$= \sum_{i=1}^{n} d_i d'_i \tag{18}$$

the classical inproduct of $\vec{d}$ and $\vec{d'}$. Also

$$|D| = \sum_{i=1}^{n} |d_i|^2 = \|\vec{d}\|^2 \tag{19}$$

$$|D'| = \sum_{i=1}^{n} |d'_i|^2 = \|\vec{d'}\|^2, \tag{20}$$

the squares of the norms of $\vec{d}$ and $\vec{d'}$. Finally

$$|D \cup D'| = \|\vec{d}\|^2 + \|\vec{d'}\|^2 - <\vec{d}, \vec{d'}> \tag{21}$$

as is readily seen. Since only (18), (19), (20) and (21) are needed in the formulae (1), (2), (3), (5), (7), (9), (11), (12) and (13), we are able to interpret all of the above measures as similarity measures for vectors. One example yields (9) :

$$\frac{|D \cap D'|}{\sqrt{|D||D'|}} = \frac{<\vec{d}, \vec{d'}>}{\|\vec{d}\|.\|\vec{d'}\|} \tag{22}$$

which is the well-known formula for $\cos(\vec{d}, \vec{d'})$, the cosine of the angle between the vectors $\vec{d}$ and $\vec{d'}$. That is why in (9), we used the term cos for the geometric mean of R and P. Note that this interpretation also shows that N (formula (11)) cannot be called a cosine function.

For the rest of this paper, however, we will limit ourselves to similarities for sets (and ordered sets).

II. General properties of similarity measures (on ordinary sets).

Let D be any real-valued function on $2^{\Omega}x2^{\Omega}$ (2^{Ω} = the set of all subsets of Ω), where D(A,B) is defined for all $A,B \subset \Omega$, $A,B \neq \varnothing$. For normalization reasons we require

$$(D_1) \quad 0 \leq D(A,B) \leq 1. \tag{23}$$

For the maximal value, D(A,B)=1, we require the highest possible similarity between A and B, i.e. $A=B\neq\varnothing$ (and vice-versa) :

$$(D_2) \quad D(A,B) = 1 \Leftrightarrow A = B \neq \varnothing \tag{24}$$

Although this is true for most of the measures discussed in the previous section, some measures do not satisfy it because they have a different purpose, e.g. measuring overlap. If we look at R, P or O_1 we see that the following property is valid

$$(D'_2) \quad D(A,B) = 1 \Leftrightarrow A \subset B \text{ or } B \subset A. \tag{24'}$$

(D'_2) is the weaker one since $(D_2) \Rightarrow (D'_2)$ obviously but, since it is a typical "overlap property" and since overlap is an important topic, we will keep both properties (D_2) and (D'_2) in our study.

For the minimal value D(A,B)=0, the situation is simpler :

$$(D_3) \quad D(A,B) = 0 \Leftrightarrow A \cap B = \varnothing. \tag{25}$$

All of the measures defined in the previous section are of the form (ϕ_1, ϕ_2 : functions)

$$D(A,B) = \frac{\phi_1(|A\cap B|)}{\phi_2(|A|,|B|,|A\cup B|)} \tag{26}$$

and satisfy

(D_4) ϕ_1 is a strictly increasing function.

Optionally, one might also require symmetry :

$$(D_5) \quad D(A,B) = D(B,A). \tag{27}$$

All measures of the previous section, except R and P, satisfy this.

In view of the above discussion we define

Definition II.1 : If D is a real-valued function on $2^{\Omega}x2^{\Omega}$ which satisfies $(D_1),(D_2),(D_3),(D_4)$, we say that D is a strong similarity measure.

Definition II.2 : If D is a real-valued function on $2^{\Omega}x2^{\Omega}$ which satisfies $(D_1),(D'_2),(D_3),(D_4)$, we say that D is a weak similarity measure.

The measures J, E (in fact all E_α), Cos, N, O_2 are strong similarity measures as is readily seen. The measures R, P, O_1 are weak similarity measures. In this article O_1 is the only symmetric weak similarity measure that is not strong, showing that $\{(D_1),(D'_2),(D_3),(D_4),(D_5)\}$ does not imply (D_2).

Note also that strong similarity measures need not to be symmetric : the measures E_α $(\alpha \neq \frac{1}{2})$ are strong similarity measures that are non-symmetric. Hence $\{(D_1),(D_2),(D_3),(D_4)\}$ does not imply (D_5).

Of course, if D is any of the above measures and if $f:[0,1]\rightarrow[0,1]$ is a strictly increasing function such that $f(0)=0$ and $f(1)=1$, then the measure $f \circ D$, defined as

$$(f \circ D)(A,B) = f(D(A,B)) \tag{28}$$

has the same properties as D. This simple remark will play a crucial role in the construction of similarity measures for ordered sets : in several cases it will only be possible to construct an OS-measure from a similarity measure D as above when using functions f as above - see the next section !.

All definitions above are given for measures $D:(A,B)\rightarrow D(A,B)$ for $A,B\neq\varnothing$. If A or B is $\varnothing$ we will define $D(A,B)=0$, for all measures D.

III. Ordered similarity measures (OS-measures)

III.1 Statement of the problem.

An ordered set is (cf. Michel (2000)) an infinite chain $C=(C_i)_{i\in\mathbb{N}}$ [where $C_i\subset\Omega$ for all $i\in\mathbb{N}$ and where $C_i\cap C_j=\varnothing$, $\forall i\neq j$] such that the elements within one C_i are unordered but for elements in different C_is we have : $d\in C_i$, $d'\in C_j$ then $d<d' \Leftrightarrow i<j$. Such an ordered set can be depicted as in Fig. 1

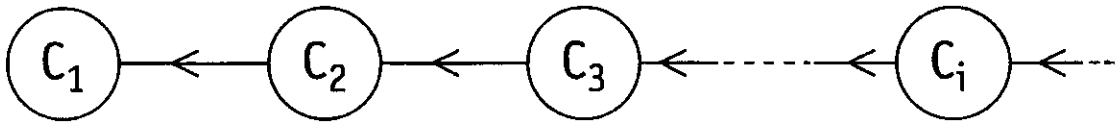

Fig. 1: Visualisation of an ordered set $C = (C_i)_{i\in\mathbb{N}}$ and where $\longleftarrow$ symbolises the order induced on the sets C_i via the order < between the documents.

Possibly $C_i=\varnothing$. This will always be the case from a certain i on, if Ω is finite. This is always so in practise but, since $|\Omega|$ can be any high number (in other words, $|\Omega|$ is unbounded), the set-up with infinite chains will prove to be very useful. Let us denote by $\mathcal{C}$ the set of all possible chains (ordered sets) as above.

This is a very general set-up, comprising the most general cases in IR : it comprises the extremes

(i) The unordered case : $C_1\neq\varnothing$, $C_i=\varnothing$, $\forall i\geq 2$

(ii) The total linear case : all C_is are singletons (or are empty).

The unordered case refers to "classical" IR (e.g. Boolean retrieval) and the total linear case refers to a ranked list of documents (as e.g. given by browsing machines in WWW). The general situation is needed e.g. in the case where one gives, say 4, key words. C_1 is then the unordered set of documents that contain all 4 of these key words. Every document in C_1 is

ordered before any document in C_2 where C_2 contains the documents that have exactly 3 of the 4 key words in their indexing. These documents, in turn, are ranked before all documents in C_3, the unordered set of documents that have exactly 2 of the 4 key words in their indexing. Documents in C_3 are ranked before all documents in C_4, the unordered set of documents that have exactly 1 of the 4 key words in their indexing. Finally, C_4 is ranked before C_5 containing the documents in $\Omega \setminus \cup_{i=1}^{4} C_i$, i.e. those with none of the 4 key words in their indexing (and $C_i=\varnothing$, $\forall i \geq 6$). One could also stop with C_4 (making $C_i=\varnothing$, $\forall i \geq 5$) since documents, not containing any of the requested key words, should not occur in a ranked output.

Ordered similarity (OS) measures compare two such chains in such a way that some "natural" properties are satisfied. We will formulate them as was done in Michel (2000), but will call them (in view of the above discussions) : strong ordered similarity measures. Their properties are : denote by Q the function, acting on couples (C,C'), $C=(C_i)_{i\in\mathbb{N}}$, $C'=(C'_j)_{j\in\mathbb{N}}$, $C,C' \in \mathcal{C}$. Q must satisfy

(Q_0) If C and C' reduce to the unordered case (cf. (i) above), Q must be a good strong similarity measure for unordered sets (i.e. satisfy (D_1), (D_2), (D_3), (D_4)).

(Q_1) $0 \leq Q(C,C') \leq 1$ for all $C,C' \in \mathcal{C}$.

(Q_2) $Q(C,C')=1 \Leftrightarrow C=C'$ and no C_i or C'_j is empty : i.e. $\forall i, \forall j, C_i \neq \varnothing, C'_j \neq \varnothing$.

(Q_3) $Q(C,C')=0 \Leftrightarrow C \cap C' = \varnothing$. Here $C \cap C' = \left(\cup_{i=1}^{\infty} C_i \right) \cap \left(\cup_{j=1}^{\infty} C'_j \right)$.

(Q_4) Let $i,j \in \mathbb{N}$, $i \neq j$. Let $C^{(i)}$, $C'^{(j)}$ be ordered sets in $\mathcal{C}$ such that $C_k \cap C'_l = \varnothing, \forall k,l \in \mathbb{N}$ except for k=i and l=j. If we let i and j vary (but not the unordered sets C_i and C'_j) then $Q(C^{(i)},C'^{(j)})$ is strictly decreasing in j>i (i fixed) and in i>j (j fixed).

(Q_5) The same as (Q_4) but now for i=j : now $Q(C^{(i)},C'^{(i)})$ strictly decreases in $i \in \mathbb{N}$.

The first four properties are clear from the unordered case ; the last two properties express the fact that sets C_i, C'_j for i,j high have a smaller impact in the comparison of C and C' than sets C_i, C'_j for i,j small.

Of course, a strong OS-measure may or may not be symmetric. If so, we have the property

(Q_6) $Q(C,C')=Q(C',C)$, $\forall C,C' \in C$.

Michel (2000) was able to construct OS-measures as follows. Let D be any similarity measure (for unordered sets). Then form

$$Q(C,C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} D(C_i,C'_j)\varphi(i,j), \tag{29}$$

where φ is a certain weighting function. The problem is to make sure that (29) satisfies the properties (Q_i) (i=0,...,5) above. The main problem here is the fact that (29) can have scores beyond 1. To avoid this Michel (2000) has put a stringent condition on the type of ordered sets that can be compared. For instance, in Michel (2000) one requires, given $C=(C_i)$ and $C'=(C'_j)$, that for every i, there exists a j such that $C_i \cap C'_j \neq \varnothing$. This condition is, however, not always valid in practise. We therefore improve Michel's method so as to be useable for comparing any two $C,C' \in C$. Point is that a solution in the form of (29) is not always possible anymore. The OS-measure (29) is therefore modified so that we obtain a generally valid solution.

III.2. General theorem on the construction of strong OS-measures.

Theorem III.2.1 : Let D be any strong similarity measure (on unordered sets). Define

$$Q(C,C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(D(C_i,C'_j))\varphi(i,j), \tag{30}$$

where f is a strictly increasing function, $f(0)=0$, $f(1)=1$, $0 \leq f \leq 1$, such that

$$\sum_{j=1}^{\infty} f(D(C_i,C'_j)) \leq 1, \ \forall i \in \mathbb{N} \tag{31}$$

$$\sum_{i=1}^{\infty} f(D(C_i,C'_j)) \leq 1, \ \forall j \in \mathbb{N} \tag{32}$$

and where φ satisfies : $\varphi>0$ and

(i) $\varphi(i,j)$ strictly decreases in $j \geq i$ (i fixed),

(ii) $\varphi(i,j)$ strictly decreases in $i \geq j$ (j fixed),

(iii) $\varphi(i,i)$ strictly decreases in i,

(iv) $\sum_{i=1}^{\infty} \varphi(i,i) = 1$,

(v) $\varphi(i,j) \leq \min(\varphi(i,i),\varphi(j,j))$, $\forall i,j \in \mathbb{N}$.

Then Q is a good strong OS-measure, i.e. satisfies (Q_0), (Q_1), (Q_2), (Q_3), (Q_4), (Q_5) (and, if D is symmetric, then Q also satisfies (Q_6) if φ is symmetric).

The proof is rather long and technical and is, therefore, given in the Appendix. This theorem will only be important if functions φ and f can be found such that (31), (32), (i)-(v) are satisfied. The construction of f is non-trivial and dependent on the concrete D (i.e. J, E, Cos, ...) and will be studied in subsection III.3. For φ, there is no problem : take e.g.

$$\varphi(i,j) = 3 \frac{1}{2^i} \frac{1}{2^j} \frac{1}{2^{|i-j|}} \tag{33}$$

This function is decreasing for high ranks i and j as well as for high differences between the ranks i and j, a natural property. Note that (33) equals the easier

$$\varphi(i,j) = \frac{3}{2^{2\max(i,j)}} \tag{34}$$

as is readily seen. It is alo readily seen that this φ satisfies all five properties (i)-(v).

That φ is a decreasing power of the ranks i and j is good and better than e.g. a linear decrease. Indeed, there should be a larger difference in comparing C_1 with C'_2 than e.g. C_{101} with C'_{102} for $C_1=C_{101}$ and $C'_2=C'_{102}$. This is also linked with the sensation law of Weber-Fechner stating that the sensation is proportional to the logarithm of the stimulus (see also Egghe (1994), Egghe and Rousseau (1990)).

We will now proceed by the construction of concrete strong OS-measures, given one of the strong similarity measures (for unordered sets) J, E, general E_α, Cos, N, O_2. For the moment it is not clear at all that a proper function f, satisfying the properties as described in the theorem,

can be constructed, for each of these strong similarity measures. Of course, any measure that yields a good f will yield a good Q, i.e. a useable strong OS-measure.

III.3. Strong OS-measures derived from strong similarity measures for ordinary sets.

III.3.1. Jaccard.

In view of theorem III.2.1 we must find a suitable f which can be used for J. Let C, $C' \in C$, $C=(C_i)_{i\in\mathbb{N}}$, $C'=(C'_j)_{j\in\mathbb{N}}$. In view of (31) and (32) we estimate $\forall i,j\in\mathbb{N}$:

$$J(C_i,C'_j) = \frac{|C_i\cap C'_j|}{|C_i\cup C'_j|}$$

$$\leq \frac{|C_i\cap C'_j|}{|C_i|} \tag{35}$$

If we fix i and let j vary, we can write

$$|C_i\cap C'_j| = \alpha_{i,j}|C_i| \tag{36}$$

where $0 \leq \alpha_{i,j} \leq 1$ and

$$\sum_{j=1}^{\infty} \alpha_{i,j} \leq 1, \tag{37}$$

since the $\alpha_{i,j}$ s denote fractions of the C'_j s in C_i (fixed). (36) in (35) yields

$$J(C_i,C'_j) \leq \alpha_{i,j} \tag{38}$$

and hence (37) gives (31) for f the identical function : f(x)=x. In the same way we have

$$J(C_i,C'_j) = \frac{|C_i\cap C'_j|}{|C_i\cup C'_j|} \leq \frac{|C_i\cap C'_j|}{|C'_j|} = \alpha'_{i,j} \tag{39}$$

where $0 \leq \alpha'_{i,j} \leq 1$ and

$$\sum_{i=1}^{\infty} \alpha'_{i,j} \leq 1 \tag{40}$$

since the $\alpha'_{i,j}$ s denote fractions of the C_is in C'_j (fixed). Hence also (32) is proved for f the identical function. Of course f(x)=x also satisfies f(0)=0, f(1)=1, $0 \leq f \leq 1$ and f strictly increasing. Conclusion :

$$Q_f(C,C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} J(C_i,C'_j))\varphi(i,j) \tag{41}$$

(with φ e.g. as in (33), (34)) is a good strong OS-measure.

III.3.2. Dice

Now, for $C,C' \in C$, we investigate

$$E(C_i,C'_j) = \frac{2|C_i \cap C'_j|}{|C_i|+|C'_j|}. \tag{42}$$

With the same definition of $\alpha_{i,j}$ as above we have :

$$\frac{2|C_i \cap C'_j|}{|C_i|+|C'_j|} \leq \frac{2|C_i \cap C'_j|}{|C_i|+|C_i \cap C'_j|}$$

$$= \frac{2\alpha_{i,j}|C_i|}{|C_i|+\alpha_{i,j}|C_i|}$$

$$= \frac{2\alpha_{i,j}}{1+\alpha_{i,j}} \tag{43}$$

We are now in search for a strictly increasing function f, $0 \leq f \leq 1$, f(0)=0, f(1)=1 such that (43) transforms into $\alpha_{i,j}$, the numbers for which (37) is true. We hence have the equation

$$f\left(\frac{2y}{1+y}\right) = y = f(x)$$

yielding

$$y = f(x) = \frac{x}{2-x} \tag{44}$$

which indeed satisfies all the requirements. Hence with this f and since it increases we have, by (37) and (43) that (31) is valid for D=E. In the same way we obtain

$$\begin{aligned} E(C_i, C'_j) &= \frac{2|C_i \cap C'_j|}{|C_i|+|C'_j|} \\ &\leq \frac{2|C_i \cap C'_j|}{|C_i \cap C'_j|+|C'_j|} \\ &= \frac{2\alpha'_{i,j}|C'_j|}{\alpha'_{i,j}|C'_j|+|C'_j|} \\ &= \frac{2\alpha'_{i,j}}{\alpha'_{i,j}+1}, \end{aligned} \tag{45}$$

with $\alpha'_{i,j}$ satisfying (40). Since (45) is the same expression as (43) we see that the same function f as in (44) will yield (32). We are now running into a surprise concerning the OS-measure derived from E in this way : indeed

$$\begin{aligned} f(E(C_i, C'_j)) &= \frac{2\frac{|C_i \cap C'_j|}{|C_i|+|C'_j|}}{2-\frac{2|C_i \cap C'_j|}{|C_i|+|C'_j|}} \\ &= \frac{|C_i \cap C'_j|}{|C_i \cup C'_j|} \\ &= J(C_i, C'_j) \end{aligned} \tag{46}$$

showing that, although we started from E, we refound the OS-measure derived from J (subsection III.3.1). In other words : $Q_E=Q_J$. Although we did not find a new OS-measure, we consider this result as a remarkable link between E and J : by the properties of f (44), we can say that E and J are very similar measures.

The most intricate case is the case of the generalized Dice indices E_α ($\alpha \neq \frac{1}{2}$). Note that these measures are not symmetrical and this causes trouble in finding the right function f. Nevertheless, we arrive at new strong OS-measures.

III.3.3 Generalized Dice.

Let $\alpha \in]0,1[\setminus\{\frac{1}{2}\}$ be fixed. We have

$$E_\alpha(C_i, C'_j) = \frac{|C_i \cap C'_j|}{\alpha|C_i| + (1-\alpha)|C'_j|}$$

$$\leq \frac{|C_i \cap C'_j|}{\alpha|C_i| + (1-\alpha)|C_i \cap C'_j|}$$

$$= \frac{\alpha_{i,j}|C_i|}{\alpha|C_i| + (1-\alpha)\alpha_{i,j}|C_i|}$$

$$= \frac{\alpha_{i,j}}{\alpha + (1-\alpha)\alpha_{i,j}}, \qquad (47)$$

with $\alpha_{i,j}$ as in (37). Transforming (47) into $\alpha_{i,j}$ gives the function

$$y = f_1(x) = \frac{\alpha x}{1-(1-\alpha)x} \qquad (48)$$

for which (31) is satisfied (for $f = f_1$). Note indeed that $0 \leq f_1 \leq 1$, $f_1(0)=0$, $f_1(1)=1$ and f_1 is strictly increasing.

Likewise we obtain

$$E_\alpha(C_i, C'_j) = \frac{\alpha'_{i,j}}{\alpha\alpha'_{i,j}+(1-\alpha)} \tag{49}$$

Transforming this into $\alpha'_{i,j}$ gives the function

$$y = f_2(x) = \frac{(1-\alpha)x}{1-\alpha x} \tag{50}$$

for which (32) is satisfied. Both functions f_1, f_2 satisfy $f_i(0)=0$, $f_i(1)=1$, $0\le f_i\le 1$, f_i strictly increasing (i=1,2) but we need <u>one</u> f in theorem III.2.1.

Hence we take

$$f = \min(f_1, f_2) \tag{51}$$

which satisfies all the requirements, notwithstanding the fact that the E_αs are non-symmetric ($\alpha \neq \frac{1}{2}$). In order to obtain workable expressions for our strong OS-measures that we obtained we need to find a concrete expression for f. Note however that

$$f_1 < f_2 \text{ iff } 0 < \alpha < \frac{1}{2}$$

$$f_2 < f_1 \text{ iff } \frac{1}{2} < \alpha < 1$$

as is readily seen. So we obtain the following good strong OS-measures. Derived from E_α with $0<\alpha<\frac{1}{2}$ we have

$$Q_{E_\alpha}(C,C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f_1(E_\alpha(C_i, C'_j))\varphi(i,j) \tag{52}$$

Here

$$f_1(E_\alpha(C_i, C'_j)) = \frac{\alpha|C_i\cap C'_j|}{\alpha|C_i|+(1-\alpha)|C'_j|-(1-\alpha)|C_i\cap C'_j|}$$

$$= \frac{\alpha|C_i\cap C'_j|}{\alpha|C_i|+(1-\alpha)|C'_j\setminus C_i|} \tag{53}$$

as is readily seen. If $\frac{1}{2}<\alpha<1$ we have

$$Q_{E_\alpha}(C,C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f_2(E_\alpha(C_i, C'_j))\varphi(i,j) \tag{54}$$

where

$$f_2(E_\alpha(C_i, C'_j)) = \frac{(1-\alpha)|C_i \cap C'_j|}{\alpha|C_i|+(1-\alpha)|C'_j|-\alpha|C_i \cap C'_j|}$$

$$= \frac{(1-\alpha)|C_i \cap C'_j|}{\alpha|C_i \setminus C'_j|+(1-\alpha)|C'_j|} \tag{55}$$

All these measures are new and non-trivial.

III.3.4 Cosine.

We will now search for the OS-measure related to Cos via (30). Now

$$\text{Cos}\,(C_i, C'_j) = \frac{|C_i \cap C'_j|}{\sqrt{|C_i||C'_j|}} \tag{56}$$

Now we have

$$\frac{|C_i \cap C'_j|}{\sqrt{|C_i||C'_j|}} \leq \frac{|C_i \cap C'_j|}{\sqrt{|C_i||C_i \cap C'_j|}}$$

$$= \frac{\alpha_{i,j}|C_i|}{\sqrt{|C_i|\alpha_{i,j}|C_i|}}$$

$$= \sqrt{\alpha_{i,j}} \tag{57}$$

,where $\alpha_{i,j}$ is as before. Now we have the requirement $\alpha_{i,j}\neq 0$ (in order to have the validity of (57)) but if $\alpha_{i,j}=0$ then Cos $(C_i,C'_j)=0$ as is $\sqrt{\alpha_{i,j}}$. So we can use (57) also for $\alpha_{i,j}=0$. Transforming (57) into $\alpha_{i,j}$, yields the function

$$y = f(x) = x^2 \tag{58}$$

obviously. By symmetry (or an analogous argument) this function also works for $\alpha'_{i,j}$. Since f(0)=0, f(1)=1, $0 \le f \le 1$ and since f increases strictly on [0,1] we have that the next measure Q is a good strong OS-measure (derived from Cos) :

$$Q_{Cos}(C, C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{|C_i \cap C_j|^2}{|C_i||C'_j|} \varphi(i,j), \tag{59}$$

a new measure.

III.3.5 The measure N.

Here

$$N(C_i, C'_j) = \sqrt{2}\ \frac{|C_i \cap C'_j|}{\sqrt{|C_i|^2 + |C'_j|^2}} \tag{60}$$

Now

$$N(C_i, C'_j) \le \sqrt{2}\ \frac{|C_i \cap C'_j|}{\sqrt{|C_i|^2 + |C_i \cap C'_j|^2}}$$

$$= \sqrt{2}\ \frac{\alpha_{i,j}|C_i|}{\sqrt{|C_i|^2 + \alpha_{i,j}{}^2 |C_i|^2}}$$

$$= \sqrt{2}\ \frac{\alpha_{i,j}}{\sqrt{1 + \alpha_{i,j}{}^2}} \tag{61}$$

with $\alpha_{i,j}$ as before. Turning (61) into $\alpha_{i,j}$ leads to the function

$$y = f(x) = \frac{x}{\sqrt{2 - x^2}} \tag{62}$$

which is strictly increasing, f(0)=0, f(1)=1, $0 \le f \le 1$ on [0,1]. The same is true for $\alpha'_{i,j}$. We have now the good strong OS-measure

$$Q_N(C, C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(N(C_i, C'_j))\varphi(i,j) \tag{63}$$

where

$$f(N(C_i, C'_j)) = \frac{|C_i \cap C'_j|}{\sqrt{|C_i|^2 + |C'_j|^2 - |C_i \cap C'_j|^2}} \tag{64}$$

as is readily seen.

III.3.6 The overlap measure O_2.

$$O_2(C_i, C'_j) = \frac{|C_i \cap C'_j|}{\max(|C_i|, |C'_j|)} \tag{65}$$

$$\le \frac{|C_i \cap C'_j|}{|C_i|}$$

$$= \alpha_{i,j}$$

and the same for $\alpha'_{i,j}$. Hence f can be taken f(x)=x. We have now the good strong OS-measure

$$Q_{O_2}(C, C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{|C_i \cap C'_j|}{\max(|C_i|, |C'_j|)} \varphi(i,j) . \tag{66}$$

Since R,P and O_1 are not strong similarity measures, they cannot be treated here as above. These weak similarity measures will be studied in another paper.

Note : As follows from the above constructions, Q_{O_2} is the largest of all the OS-measures that are constructed here. Indeed, denoting Q_D for any of the $Q_J=Q_E$, Q_{E_α}, Q_{Cos}, Q_N, we have that

$$f(D((C_i, C'_j)) \leq \min(\alpha_{i,j},\alpha'_{i,j})$$

$$= \frac{|C_i \cap C'_j|}{\max(|C_i|,|C'_j|)}, \tag{67}$$

hence $Q_D \leq Q_{O_2}$, for all the above D. Note the difference with (14).

The next section investigates some of these measures in a practical contexts.

IV. Experimentation

IV.1 Presentation of the context

The experimentation's aim is to compare different OS-measures. In order to test it (as in Michel (2000)), we have constructed a corpus of 550 queries and put them to the personalized filtering information retrieval system, Profil-Doc, developed by the laboratory RECODOC - see Michel (1999). Profil-Doc has the particularity of making a filtering of information regarding the profile of the user (Lainé-Cruzel (1996)). SPIRIT©[2], the motor used to extract documents from the plain text data base, uses weights to present answer documents into classes, theses classes being ranked in order of relevance to the user (see Fluhr (1997)).

Three different users profiles (called T1, T2 and T3) are tested with the 550 queries. In order to measure profiles' effectiveness we compare, for each query, the personalized answer with a

[2] SPIRIT (Syntactic and Probabilistic Indexing and Retrieval Information System) is a commercial product of T.GID. Searches about SPIRIT are made according to the CEA-DIST (Atomic Energy Commission - Sientific and Technique Information Direction) - http://www.dist.cea.fr/

neutral one, i.e. given by the system without any filtering. In the four following figures we can see the number of documents (on the Y-axis) given by the system for each answer (number of answer on the X-axis). For each curve, values are presented by decreasing number of documents for T1. So query numbers cannot be directly compared.

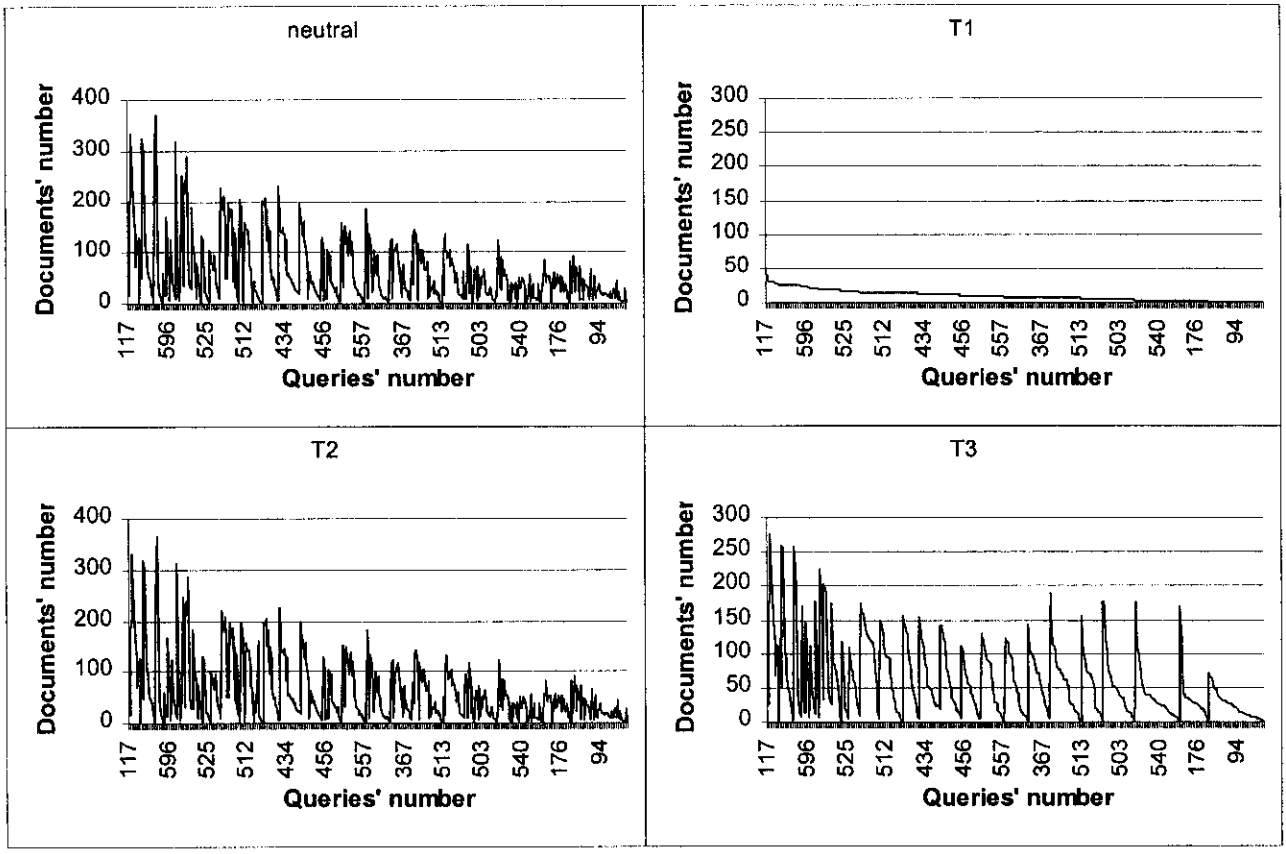

Figure 2 : Number of documents per query

We can see that the neutral process, T2 filtering and T3 filtering give the same variety of answers relating to the number of documents, indeed it varies in a regular way from 1 to respectively 371, 366 and 276. T1 filtering gives more specialized answers (indeed none of them have more than 40 documents) but as before, we can observe a regular distribution. So the number of documents per answer is not a bias in the experimentation.

OS-measures, presented in the following section, will be used to calculate, query per query, the similarity degree between neutral and personalized answers.

IV.2 Tested OS-measures

We choose to compare the Jaccard and Cosine strong OS-measures defined previously in equations (41) and (59) by (we used finite sums since this is always the case in practise) :

$$Q_J(C, C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} J(C_i, C'_j)\varphi(i,j)$$

$$Q_{Cos}(C, C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{|C_i \cap C'_j|^2}{|C_i||C'_j|}\varphi(i,j)$$

As we have said before, it is possible to choose for φ any function verifying the conditions (i), (ii), (iii), (iv), (v) of theorem III.2.1.

We have shown in Michel (2000) that the linear function $\varphi_l(i,j)$ defined as follows is correct.

$$\varphi_l(i,j):[1,m]x[1,m']\to\mathbb{R}^+;\varphi_l(i,j)=\delta^{mo}(i(|i-j|+1))x\delta^{mo}(j(|i-j|+1)) \quad (68)$$

where $\delta^{mo}:[1,m_0^2]\to\mathbb{R}^+$, $m_o=\max(m,m')$

$$\delta^{m_0}(n) = \sqrt{\frac{6m_0^3}{6m_0^4-6m_0^3+8m_0^2-3m_0+1}}\left(1-\frac{n-1}{m_0^2}\right) \quad (69)$$

We have seen in equation (33), that a power function like $\varphi_l(i,j)=\frac{3}{2^i 2^j 2^{|i-j|}}$ is good too. For normalization reasons with finite sum, we now apply a normalization constant equalling $\frac{4^{m_0}}{4^{m_0}-1}$. So, the second weight function is defined by :

$$\varphi_p:[1,m]x[1,m']\to\mathbb{R}^+;\varphi_p(i,j)= \frac{4^{m_0}}{4^{m_0}-1}\frac{3}{2^i 2^j 2^{|i-j|}}. \quad (70)$$

We hence have now four strong concrete OS-measures defined as follows

$$J^{\Omega}_l(C, C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} J(C_i, C'_j)\varphi_l(i,j) \quad (71)$$

$$J^{\Omega}_{p}(C, C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} J(C_i,C'_j)\varphi_p(i,j) \quad (72)$$

$$Cos^{\Omega}_{I}(C, C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{|C_i \cap C'_j|^2}{|C_i||C'_j|}\varphi_I(i,j) \quad (73)$$

$$Cos^{\Omega}_{p}(C, C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{|C_i \cap C'_j|^2}{|C_i||C'_j|}\varphi_p(i,j) \quad (74)$$

The measure Q_{Rec} defined in (75) is derived from the Recall.

$$Q_{Rec}(C, C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{|C_i \cap C'_j|}{|C_i|}\varphi(i,j) \quad (75)$$

Recall has been found to be a weak similarity measure, hence not suited in our framework. But it is sufficient for a "simple" OS-measure as defined in Michel (2000). We choose to test it only for experimentation purpose. The corresponding formulae are :

$$R^{\Omega}_{I}(C, C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} R(C_i,C'_j)\varphi_I(i,j) \quad (76)$$

$$R^{\Omega}_{p}(C, C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} R(C_i,C'_j)\varphi_p(i,j) \quad (77)$$

IV.3 Analysis

We will make three types of comparisons in order to estimate

- the impact of the weight function φ,
- the impact of the similarity indicators (J, Cos and R) and
- the impact of the query type (specialized or general queries).

The impact of the weight function φ is estimated by comparing J^{Ω}_{I} and Cos^{Ω}_{I} with their corresponding means J_{m_0} and Cos_{m_0} defined as :

$$J_{m_0} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m'} J(C_i,C'_j)}{m_0}$$

and

$$Cos_{m_0} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m'} Cos(C_i,C'_j)}{m_0}.$$

The means are considered as measures with neutral weight function. In order to make the curves readable, we have ranked the queries according to increasing values of J^{Ω}_1 and Cos^{Ω}_1 (figures 3,4,5,6,7,8).

The impact of the similarity indicator J, Cos and R is estimated by comparing J^{Ω}_1, Cos^{Ω}_1, R^{Ω}_1, and J^{Ω}_p, Cos^{Ω}_p, R^{Ω}_p. In order to make the curves readable, we have ranked the queries according to the increasing values of J^{Ω}_p (figures 9,10,11).

The impact of the query type (specialized or general queries) is estimated by the shape of the curve when we rank the results of J^{Ω}_1 in function of the increasing number of documents per answer (figure 12), the increasing number of classes per answer (figure 13) and the increasing number of documents per class per answer (figure 14).

IV.4 Results

IV.4.1 Impact of weight function

For each user's interrogation context T1, T2 and T3, we compare the OS J^{Ω}_1 and Cos^{Ω}_1 with their corresponding non ordered rank measure J_{m_0} and Cos_{m_0}. We have chosen not to present the comparison of (J^{Ω}_p, J_{m_0}) and (Cos^{Ω}_p, Cos_{m_0}) because we compare (J^{Ω}_1, Cos^{Ω}_1) and (J^{Ω}_p, Cos^{Ω}_p) in the following section. In all figures, the queries on the X-axis are represented according to the increasing values of J^{Ω}_1 or Cos^{Ω}_1.

Bold curves relate to the OS linear measures (J^{Ω}_1 and Cos^{Ω}_1). Thin curves relate to the corresponding mean (J_{m_0}, Cos_{m_0}). We notice very different shapes of curves between the T1, T2 and T3 contexts. Indeed, when we rank the results of T1 by increasing values of J^{Ω}_1 we can observe an exponential type function. The same operation in T2 results in a logarithmic type function and on T3 results in an S-shaped function. These differences show that the systems to be compared give really different and personalized results.

Globally, the bold and thin curves have the same shape. In all cases, generally, the thin curve is above the bold one. The interpretation is the same as in Michel (2000) : *"The J_{m_0} curve*

slightly above the OS curve is a consequence of the choice of the decreasing function δ^{m_0}. *Punctually, we can observe values of* J_{m_0} *below the curve. This phenomenon is explained by the calculation of the average of the index of Jaccard on* m_0, *the maximum of the classes of the compared sets. If m and m' are the numbers of classes of each compared answer and if there is a strong difference between m and m': then* J_{m_0} *could be lower than the corresponding OS value. The further the curves* J_{m_0} *and OS will move away, the more important the filtering will be in the scheduling of the documents".*

By comparing the figures presented on the same line (figures 3 and 4, figures 5 and 6, figures 7 and 8), we can remark that the values of J and Cos look like very similar. There are 3 zones:

- Zone 1 (called head of the curve) is when J_{m_0} and J^{Ω}_1 are identical to 0. This phenomenon must have been visisble in the curves 3 and 6. It didn't because for technical reasons we have had to suppress some queries.
- Zone 2 (called heart of curve) is when J_{m_0} (respectively Cos_{m_0}) takes extremely variable values compared to J^{Ω}_1 (respectively Cos^{Ω}_1)
- Zone 3 (called tail of the curve) is when J_{m_0} and J^{Ω}_1 (or Cos_{m_0} and Cos^{Ω}_1) are identical to 1.

We can observe that the tail of curve is unchanged from the Jaccard case (J^{Ω}_1 , J_{m_0}) to the cosine case (Cos^{Ω}_1, Cos_{m_0}). On the contrary, in the heart of the curve (zone 2) the thin curve of Cos_{m_0} has much more amplitude than J_{m_0}. We can say that, **without any weight function, the cosine measure seems to be less stable than the Jaccard**.

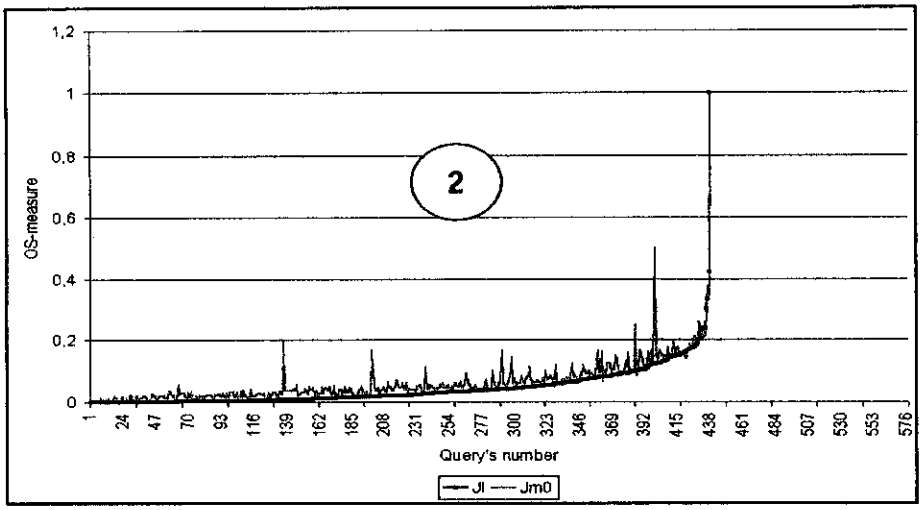

Figure 3 : T1 - J^{Ω}_{I}, J_{m0}

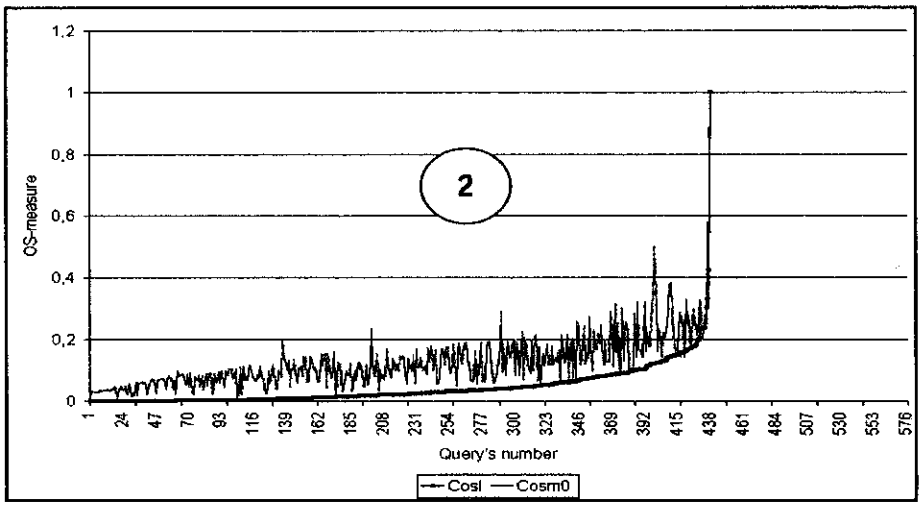

Figure 4 : T1 - Cos^{Ω}_{I}, Cos_{m0}

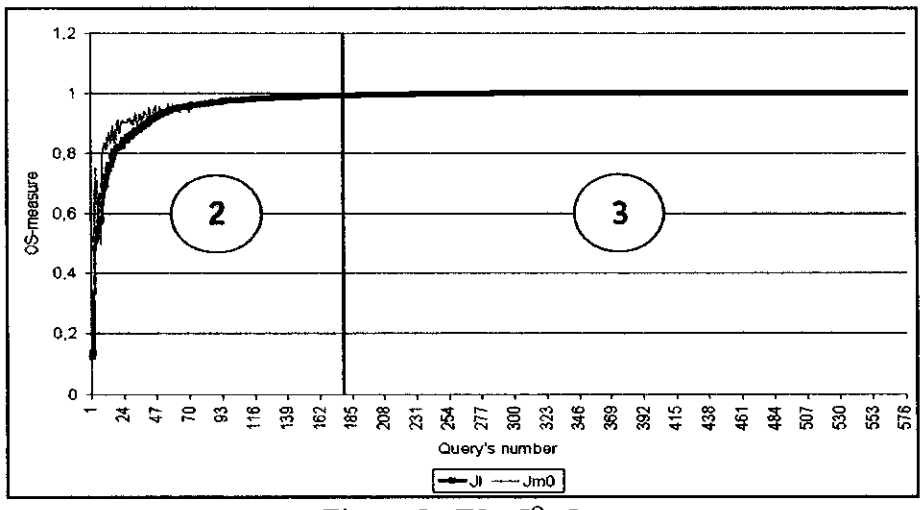

Figure 5 : T2 - J^{Ω}_{I}, J_{m0}

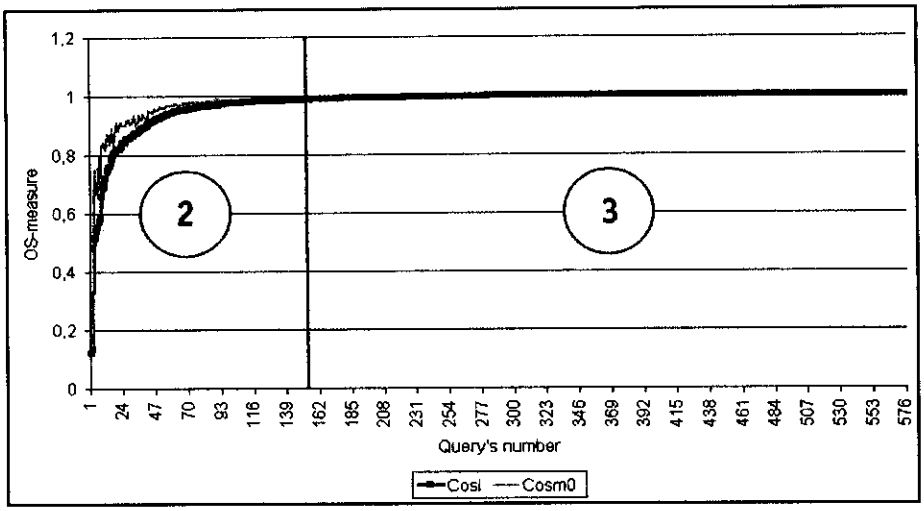

Figure 6 : T2 - Cos^{Ω}_{l}, Cos_{m0}

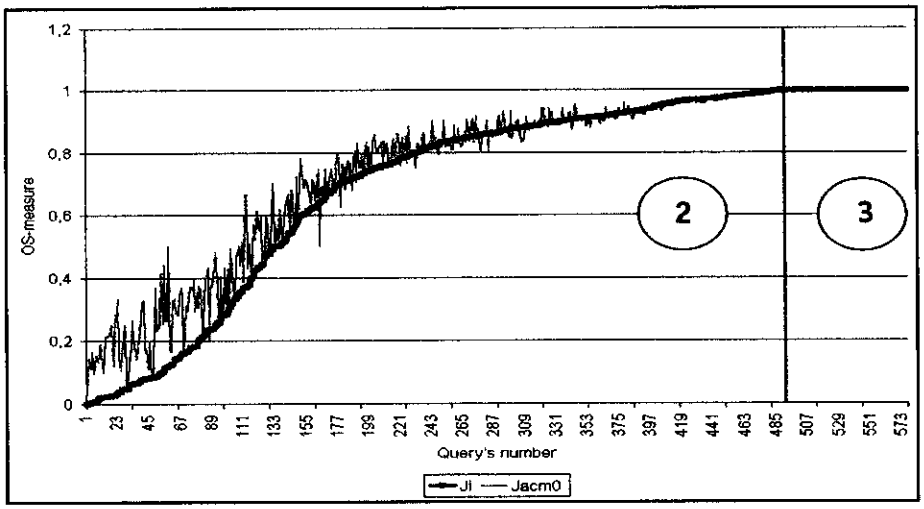

Figure 7 : T3 - J^{Ω}_{l}, J_{m0}

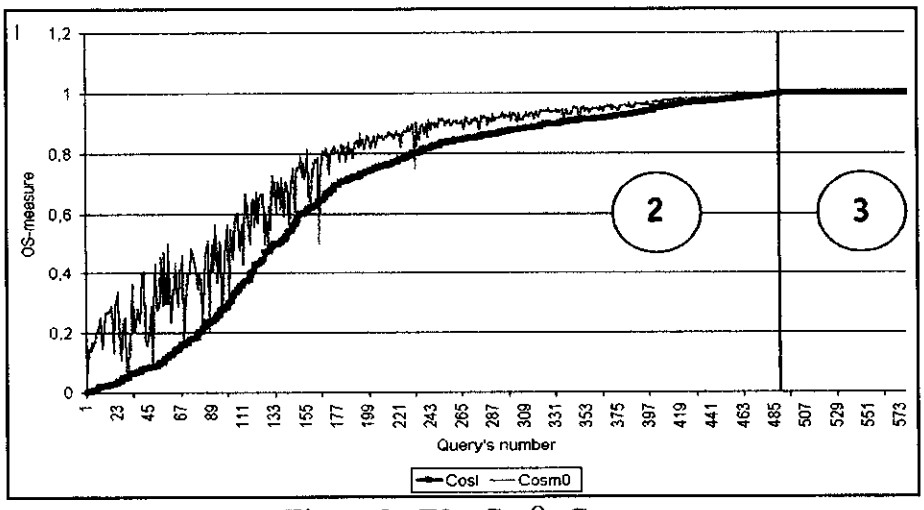

Figure 8 : T3 - Cos^{Ω}_{l}, Cos_{m0}

IV.4.2 Impact of the classical similarity indicator

In the three following figures we compare, for T1, T2 and T3, the OS J^{Ω}_{l}, Cos^{Ω}_{l}, R^{Ω}_{l} and J^{Ω}_{p}, Cos^{Ω}_{p}, R^{Ω}_{p}. In order to make the curves readable, we have ranked the queries according to increasing values of J^{Ω}_{p} (figure 9,10,11).

In the three figures we can remark that for a given weight function and irrespective of the indicator used (Cosine, Jaccard or Recall) the results of OS measures are strictly identical for each of the 550 queries. Indeed, we have $J^{\Omega}_{l} \approx Cos^{\Omega}_{l} \approx R^{\Omega}_{l}$ (represented by the thin curve) and $J^{\Omega}_{p} \approx Cos^{\Omega}_{p} \approx R^{\Omega}_{p}$ (bold curve).

This very remarkable result shows that the weight function suppresses the initial effect of the Jaccard, Cosine or Recall measure.

Secondly, we can remark that, in each figure, thin and bold curves' shapes are identical. Indeed, as above, with T1 we can observe an exponential function (figure 9), with T2 there is a logarithmic function (figure 10) and with T3 we have an S-shaped function (figure 11). This regularity shows that the linear and power indicator act in the same way but not on the same degree.

Finally we can say that in each system, the thin curve is always above the bold one in the beginning, goes through it in a particular point and remains below until the end of the values. When the answers are very different (similarity near 0), the power weight is more precise than the linear one. On the contrary, when the answers are very similar, the linear function is more precise. We did not find any interpretation of the intersection point.

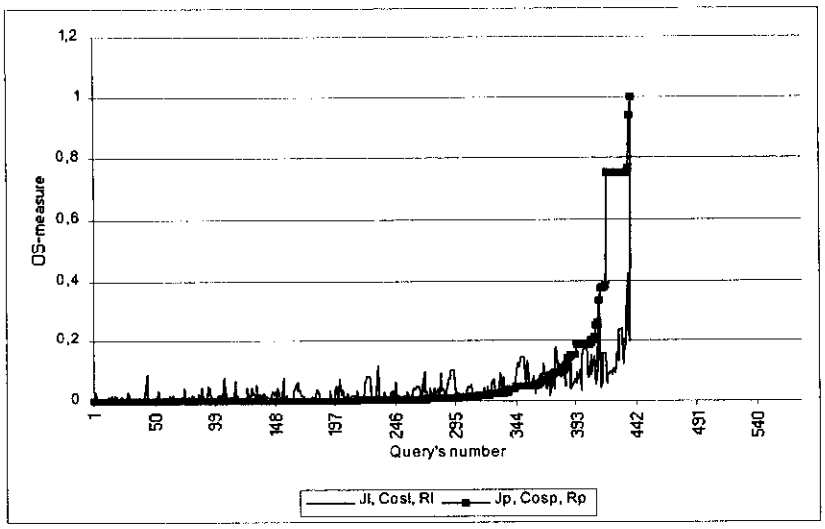

Figure 9 :T1 - J^{Ω}_{l}, Cos^{Ω}_{l}, R^{Ω}_{l} and J^{Ω}_{p}, Cos^{Ω}_{p}, R^{Ω}_{p}

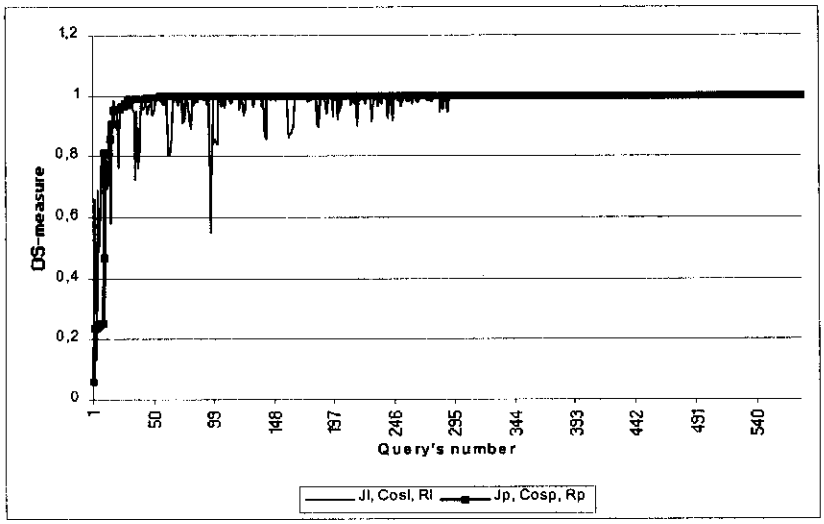

Figure 10 : T2 - J^{Ω}_{l}, Cos^{Ω}_{l}, R^{Ω}_{l} and J^{Ω}_{p}, Cos^{Ω}_{p}, R^{Ω}_{p}

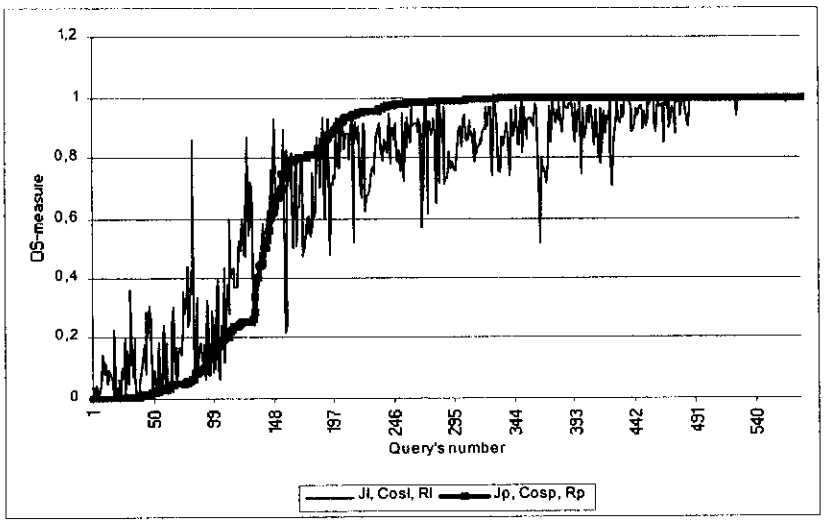

Figure 11: T3 - J^{Ω}_{l}, Cos^{Ω}_{l}, R^{Ω}_{l} and J^{Ω}_{p}, Cos^{Ω}_{p}, R^{Ω}_{p}

IV.4.3 Impact of the query type

We have some quantitative information about the answers, for example the number of documents they have or the number of classes they have.

Let us suppose first that the query type is a function of the number of documents that the corresponding answer gives : a lot of document means that the query is very general ; only few documents means that the query is specialized. In order to observe if this fact as an impact on the similarity measure, we rank the results of J^{Ω}_{l} (thin curve) and J^{Ω}_{p} (bold curve) in function of the increasing number of documents per answer. In figure 12, the curves are drawn with results of user's T2. We cannot observe any regularity. Some queries have very particular results, giving abrupt decreasing values. There is nothing particular characterizing these queries. For different reasons, linked with the particular profile T2 of the user, the personal answers are very different to the neutral ones. In the case of T1 and T3 (not represented with figures) the same phenomenon is observed for other queries.

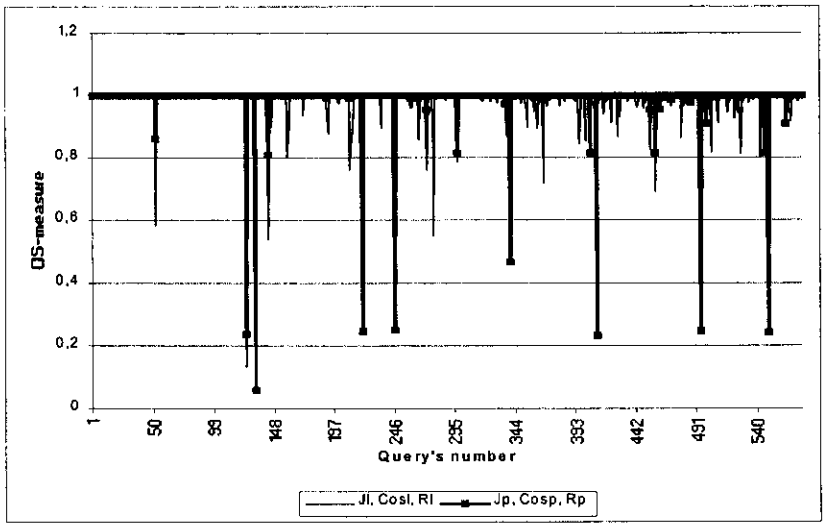

Figure 12 : T2 - J^{Ω}_{l}, and J^{Ω}_{p}, ranking by increasing number of documents per answer

We know that in the answers, the documents are grouped into classes. The more the queries have terms the more the answers may have classes. So, if we suppose that general queries have few terms, answers with little number of classes must be supposed to be more general. In order to test if the number of classes has an impact on the OS, we have ranked the answers in function of the number of classes (figure 13). But, as before, there is no regularity.

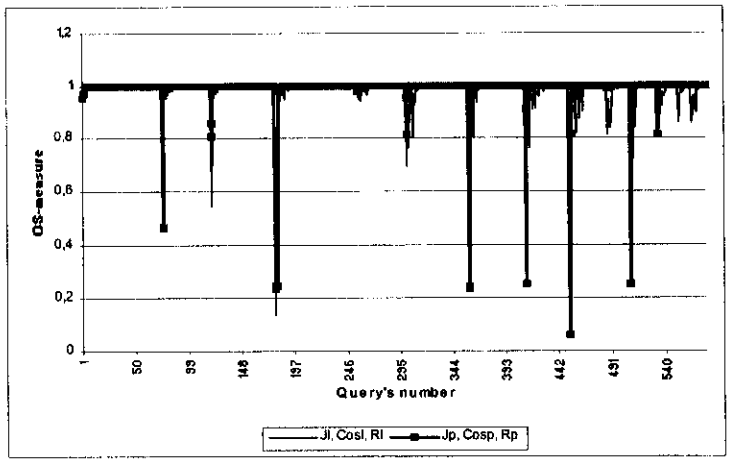

Figure 13 : T2 - J^{Ω}_{I}, and J^{Ω}_{p}, ranking by increasing number of class per answers

The last and final test is to rank the queries in function of the number of documents per class. But in this case too, no regularity is observed (figure 14).

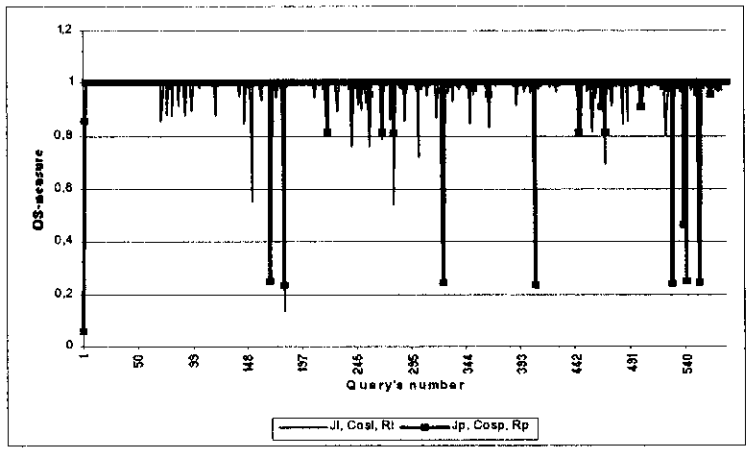

Figure 14 : T2 - J^{Ω}_{I}, and J^{Ω}_{p}, ranking by increasing number of document/class per answers

V. Conclusion

In Michel (2000) we have presented some first results on OS-measures. The present work completes it by presenting a more general treatment of OS-measures. Indeed, in section II we make distinction between two types of OS-measures : strong and weak ones. In section III, a method is proposed to construct strong OS-measures. It is usable in more general contexts

than the one presented in Michel (2000) because it works for comparing any two ordered sets whereas Michel (2000) needs some restrictive compatibility condition on them. We propose in section III concrete good strong OS-measures derived from strong similarity measures of Jaccard, Dice (from which we refind the strong OS-measures derived from Jaccard), Generalized Dice, Cosine, N, Overlap O_2. We remark that we did not have taken into account the case of R, P and O_1 because they are not strong similarity measures. These weak similarity measures will be studied in another paper.

We choose to test two strong OS-measures derived from Jaccard and Cosine, and one simple OS-measure derived from Recall. The construction of concrete OS-measures is made with a linear and a power weight function. We made no hypothesis upon the similarity indicator. We made one hypothesis on the weight function : according to the law of Weber-Fechner, a power weight function is supposed to be better than a linear one to reflect the rank difference between ordered sets.

First results show that without any weight function the Jaccard indicator seems to be more stable than the Cosine one. This result is maybe linked with the fact that Spirit©, the software use in the experimentation, is a weighted Boolean system. The second very remarkable result is that the weight function (linear or power) suppresses, for both strong and simple OS results, the initial effect of the Jaccard, Cosine or Recall indicator. Comparative tests on the linear and power weight function show that they act in the same way but not on the same degree : there is not one better than the other. Indeed, when the answers are very different (similarity near 0), the power weight is more precise than the linear one. On the contrary, when the answers are very similar, the linear function is more precise. Nothing has been remarked on a possible link between the type of query (general or specialized query) and the OS characteristics.

Some complementary experimentations with the Overlap measure O_2 and the generalized Dice could be interesting in order to know if these phenomenon are general to all strong OS-measures. In this case, it would be interesting to know if it is really important to find some "weakest" weight function, or if (and in this case why) in the ordered sets cases, the rank is the more important criterion. These will be studied in another paper.

References

Boyce, B.R., Meadow, C.T., & Kraft, D.H. (1995). Measurement in information science. Academic Press, New York.

Egghe, L. (1994). A theory of continuous rates and applications to the theory of growth and obsolescence rates. Information Processing and Management, 30(2), 279-292.

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

Fluhr, C. (1997). SPIRIT.W3 : A distributed Cross-Lingual Indexing and Search Engine. Proceedings of the INET 97 "The Seventh Annual Conference of the Internet Society". June 24-27 1997. Kuala Lumpur, Malaysia.

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

Lainé-Cruzel, S., Lafouge, T., Lardy, J.P. & Ben Abdallah, N. (1996). Improving information retrieval by combining user profile and document segmentation. Information Processing and Management, 32(3), 305-315.

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

Michel, C. (1999). Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes. Réalisation et evaluation d'un prototype de système de recherche d'information avec filtre selon les profils des utilisateurs. Ph.D. Thesis. University Lyon II. 6 January 1999. 322p.

Michel, C. (2000). Ordered similarity measures taking into account the rank of documents. Information Processing and Management, to appear.

Salton, G. & Mc Gill, M.J. (1987). Introduction to modern Information Retrieval. Mc Graw-Hill, New York.

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

Van Rijsbergen, C.J. (1979). Information Retrieval, 2nd Edition. Butterworths, London.

Appendix

Proof of theorem III.2.1

(Q_0)

If $C=(C_1,\varnothing,\ldots,\varnothing,\ldots)$ and $C'=(C'_1,\varnothing,\ldots,\varnothing,\ldots)$, symbolising the unordered case, then (30) reduces to

$$Q(C,C') = f(D(C_1,C_1')),$$

hence the measure of $f \circ D$. We noted already (in section II) that $f \circ D$ is a good strong similarity measure if D is, because of the properties of f. We leave the easy proof to the reader.

(Q_1)

for all $C, C' \in \mathcal{C}$:

$$0 \leq Q(C,C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(D(C_i,C'_j))\, \varphi(i,j)$$

$$\leq \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(D(C_i,C'_j))\, \varphi(i,i)$$

$$\leq \sum_{i=1}^{\infty} \varphi(i,i) = 1,$$

using (31), (iv) and (v). Note that, instead of (31), we could have used (32) as well. For the proof of (Q_2) we will, however, need <u>both</u> (31) and (32).

(Q_2)

Let $Q(C,C')=1$ for $C=(C_i)_{i\in N} \in \mathcal{C}$ and $C'=(C'_j)_{j\in N} \in \mathcal{C}$. Hence

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(D(C_i,C'_j))\, \varphi(i,j) = 1 \qquad \text{(A1)}$$

Suppose $\exists j_0 > i_0$ such that $f(D(C_{i_0}, C'_{j_0})) \neq 0$.

Then, by (i)

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(D(C_i, C'_j))\, \varphi(i,j) < \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(D(C_i, C'_j))\, \varphi(i,i)$$

$$\leq \sum_{i=1}^{\infty} \varphi(i,i) = 1,$$

by (31) and (iv), contradicting (A1). Suppose $\exists i_0 > j_0$ such that $f(D(C_{i_0}, C'_{j_0})) \neq 0$.

Then, by (ii)

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(D(C_i, C'_j))\, \varphi(i,j) < \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(D(C_i, C'_j))\, \varphi(j,j)$$

$$= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} f(D(C_i, C'_j))\, \varphi(j,j)$$

$$\leq \sum_{j=1}^{\infty} \varphi(j,j) = 1,$$

by (32) and (iv), again contradicting (A1).Hence

$$Q(C,C') = \sum_{i=1}^{\infty} f(D(C_i, C'_i))\, \varphi(i,i) = 1. \qquad \text{(A2)}$$

If there exists an $i \in \mathbb{N}$ such that $f(D(C_i, C'_i)) < 1$, then by (iv) and the fact that $\varphi > 0$:

$$Q(C,C') < \sum_{i=1}^{\infty} \varphi(i,i) = 1,$$

again a contradiction. Hence $f(D(C_i, C'_i)) = 1$, $\forall i \in \mathbb{N}$. By the proporties of f we have $D(C_i, C'_i) = 1$ and since D is a strong similarity measure (hence satisfying (D_2)) we have that $C_i = C'_i$, $\forall i \in \mathbb{N}$. Hence $C = C'$. Conversely, if $C = C'$, we have that $C_i \cap C'_j = \varnothing$ $\forall i,j$, $i \neq j$ and $C_i = C'_i$, $\forall i \in \mathbb{N}$. Hence (by the properties of D and f) :

$$Q(C,C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(D(C_i, C'_j))\, \varphi(i,j)$$

$$= \sum_{i=1}^{\infty} f(D(C_i, C'_i))\, \varphi(i,i)$$

$$= \sum_{i=1}^{\infty} f(D(C_i,C_i))\ \varphi(i,i)$$

$$= \sum_{i=1}^{\infty} \varphi(i,i) = 1.$$

(Q_3)

$$Q(C,C') = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(D(C_i,C'_j))\ \varphi(i,j) = 0$$

iff $f(D(C_i,C'_j)) = 0$, $\forall i,j \in \mathbb{N}$ (since $\varphi>0$). By the fact that f strictly decreases and $f(0)=0$ we see that $D(C_i,C'_j) = 0$. Hence (D_3) implies $C_i \cap C'_j = \varnothing$ $\forall i,j \in \mathbb{N}$. Hence $\forall i \in \mathbb{N} : C_i \cap \left(\underset{j=1}{\overset{\infty}{\cup}} C'_j \right) = \varnothing$, hence $\left(\underset{i=1}{\overset{\infty}{\cup}} C_i \right) \cap \left(\underset{j=1}{\overset{\infty}{\cup}} C'_j \right) = \varnothing$, hence $C \cap C' = \varnothing$.

(Q_4)

For $C^{(i)}, C'^{(j)} \in C$ as in (Q_4) we have

$$Q(C^{(i)},C'^{(j)}) = f(D(C_i,C'_j))\ \varphi(i,j). \qquad (A3)$$

As given, $f(D(C_i,C'_j))$ is constant in i and j (the sets remain the same ; only their ranks are variable). Hence (Q_4) follows from (i) and (ii).

(Q_5)

For $C^{(i)}, C'^{(i)} \in C$ as above we have

$$Q(C^{(i)},C'^{(i)}) = f(D(C_i,C'_i))\ \varphi(i,i).$$

Again, $f(D(C_i,C'_i))$ is independent of i. Hence (Q_5) follows from (iii).

Of course, if D is symmetric and φ is too, then (30) implies that Q is symmetric too. □