Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques

EGGHE, Leo & Michel, Ch. (2003) Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques. In: Information Processing & Management, 39(5). p. 771-807.

DOI: 10.1016/S0306-4573(02)00027-4
Handle: http://hdl.handle.net/1942/778

CONSTRUCTION OF WEAK AND STRONG SIMILARITY MEASURES FOR ORDERED SETS OF DOCUMENTS USING FUZZY SET TECHNIQUES

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

and

C. Michel, CEM-GRESIC, MSHA, D.U. Bordeaux III, Esplanade des Antilles, F-33607, Pessac Cedex, France[1]
Christine.Michel@montaigne.u-bordeaux.fr

ABSTRACT

Ordered sets of documents are encountered more and more in information distribution systems, such as information retrieval systems (IRS). Classical similarity measures for ordinary sets of documents hence need to be extended to these ordered sets. This is done in this paper using fuzzy set techniques. First a general similarity measure is developed which contains the classical strong similarity measures such as Jaccard, Dice, Cosine and which contains the classical weak similarity measures such as Recall and Precision.

Key Words : similarity measure, ordered set, fuzzy

---

[1] Permanent address.

Then these measures are extended to comparing fuzzy sets of documents. Measuring the similarity for ordered sets of documents is a special case of this, where, the higher the rank of a document, the lower its weight is in the fuzzy set. Concrete forms of these similarity measures are presented. All these measures are new and the ones for the weak similarity measures are the first of this kind (other strong similarity measures have been given in a previous paper by Egghe and Michel).

Some of these measures are then tested in the IR-system Profil-Doc. The engine SPIRIT© extracts ranked documents sets in 3 different contexts, each for 600 request. The practical useability of the OS-measures is then discussed based on these experiments.

I. Introduction

The following similarity measures are well-known measures to compare sets of documents A,B (subsets of a documentary system, Ω, the universe).

Jaccard's index J

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

Dice's index D

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

Cosine function Cos

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

The measure N

$$N(A,B) = \sqrt{2}\frac{|A \cap B|}{\sqrt{|A|^2 + |B|^2}} \tag{4}$$

Overlap measures O_1 and O_2

$$O_1(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} \tag{5}$$

$$O_2(A,B) = \frac{|A \cap B|}{\max(|A|,|B|)} \tag{6}$$

Recall R and Precision P

$$R(A,B) = \frac{|A \cap B|}{|B|} \tag{7}$$

$$P(A,B) = \frac{|A \cap B|}{|A|} \tag{8}$$

Of course here one must add that A=ret, the set of retrieved documents and B=rel, the set of relevant documents.

Most of these measures are classical and well-known. For the readers who wish a longer introduction on these measures we refer to van Rijsbergen (1979), Salton and Mc Gill (1987), Boyce, Meadow and Kraft (1995), Tague-Sutcliffe (1995), Grossman and Frieder (1998), Losee (1998) or Egghe and Michel (2002).

All of the above measures are of the form

$$F(A,B) = \frac{\psi_1(|A \cap B|)}{\psi_2(|A|,|B|,|A \cup B|)} \tag{9}$$

where ψ_1 is a strictly increasing function and ψ_2 is an increasing function of 3 variables.

Let us consider the following general properties for a general measure F : for all $A,B \subset \Omega$:

(F_1) $0 \leq F(A,B) \leq 1$ (10)

(F_2) $F(A,B) = 1 \Leftrightarrow A = B$ (11)

(F_3) $F(A,B) = 0 \Leftrightarrow A \cap B = \emptyset$ (12)

(F_4) If the denominator of F is constant then F is strictly increasing with $|A \cap B|$.

The measures J, D, Cos, N, O_2 satisfy these four requirements and are called strong similarity measures. If we replace (F_2) by the weaker

(F'_2) $F(A,B) = 1 \Rightarrow A \subset B$ or $B \subset A$ (13)

then we have that O_1, R and P satisfy (F_1), (F'_2), (F_3), (F_4) and are therefore called weak similarity measures. Not that they are less important than the strong ones : the applications of e.g. R and P in IR are well-known ! Note that the reverse of (13) is OK for O_1 but not for R or P. Indeed suppose $A \subset B$ (and not $B \subset A$). Then $P(A,B)=1$ but $R(A,B)<1$. The same can be said when $A \supset B$ (and not $B \supset A$) with R and P reversed. That is why we kept $\Rightarrow$ in (13).

So far for unordered sets (i.e. ordinary sets) of documents.

In Michel (2000) and Egghe and Michel (2002) one considers ordered sets of documents as e.g. the output of an IR query. The most general case can be depicted as in Fig. 1 where one has a "chain" of unordered disjoint sets $C_i \subset \Omega, \forall i \in \mathbb{N}$ but such that every document in C_{i-1} is before (i.e. has a lower rank) every document in C_i, for every i. In terms of IR, the documents in C_{i-1} are retrieved before (or read before) the ones of C_i, for every i. Of course $C_i = \emptyset$ will always be the case, in practise, from a certain i on (since Ω is finite in practise).

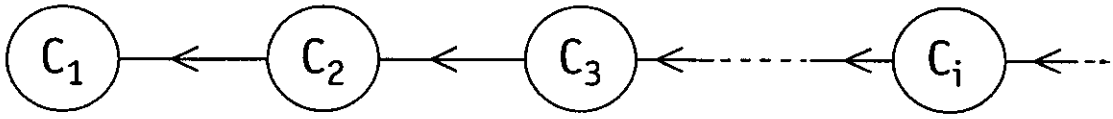

Fig. 1 An ordered set $C=(C_i)_{i \in \mathbb{N}}$ where --<-- symbolises the order induced on the sets C_i via the order between the documents.

We work here with a very general set-up, comprising the IR extremes :

(a) the unordered case : $C_1 \neq \emptyset$, $C_i = \emptyset$, $\forall i \geq 2$

(b) the total linear case : $\forall i: |C_i| = 1$ or $C_i = \emptyset$

As explained in Egghe and Michel (2002), case (a) represents "classical" IR (e.g. Boolean retrieval) and the total linear case (b) refers to a ranked list of documents as e.g. given by browsing machines in WWW. The very general situation is needed in Boolean retrieval with OR-ed key words (say N in number) and where C_i is the set of documents that contain N-i+1 of the N given key words (i=1,...,N) (all the other C_i are empty) : here there is no natural order within each C_i but, as in our model, every document in C_{i-1} is before (is potentially more important) than every document in C_i, since the former contains more requested key words than the latter.

In Egghe and Michel (2002) we used the above model to construct ordered similarity (OS) measures, i.e. measures that can compare every two such chains $C=(C_i)_{i\in\mathbb{N}}$ and $C'=(C'_j)_{j\in\mathbb{N}}$. They were derived from existing good similarity measures for unordered sets, applied to each C_i and C'_j and then "combining" everything together by summing over each i and j in $\mathbb{N}$. Here the sets C_i and C'_j on the higher ranks receive a lower weight than the ones on the lower ranks. We were then able to construct good <u>strong</u> OS measures, i.e. measures Q on chains C and C' such that several "natural" properties are satisfied. One of them is that $Q(C,C')=1$ if and only if $C=C'$ (as a chain, i.e. $C_i=C'_i$, $\forall i\in\mathbb{N}$) and no C_i or C'_j is empty. This boils down to the property (F_2) in the case of unordered sets. It is obvious then that the method used in Egghe and Michel (2002) only works when strong similarity measures are concerned.

Since we do not have a natural analogue of (F'_2) for chains (i.e. what is the meaning of $C \subset C'$?), we are, in this paper considering the chains as fuzzy sets, where inclusion ($\subset$) is a clearly defined thing. This will yield a new model for good OS measures. Our technique will work for weak as well as strong measures. Note that we will here present good weak OS measures for the first time.

The next section deals with a classification and a generalization of strong and weak similarity measures. These generalized measures enable us to present one fuzzy model for weak and strong OS measures, which is then applicable to all the measures encountered so far.

The third section presents a general theory of weak and strong similarity measures on fuzzy sets. It is new in itself and more general than the application given in section four to ordered sets. There we interpret our "chains" (as in Fig. 1) in the sense of fuzzy sets and give concrete formulae in this case. In this way new good OS measures (weak and strong ones) are created, hence producing good weak OS measures for the first time.

The last section reuses the IR-system Profil-Doc and the engine SPIRIT©, where ranked document sets for 600 requests in 3 different contexts are given. On these sets some of the measures, obtained in this paper, are calculated and based on graphical representations of the values of these OS-measures, qualitative conclusions on these OS-measures are given.

II. Generalized forms of weak and strong similarity measures on ordinary sets.

In the introduction, we mentioned already that the measure F, defined on $2^\Omega \times 2^\Omega$ (2^Ω = the set of all subsets of Ω) as

$$F(A,B) = \frac{\psi_1(|A \cap B|)}{\psi_2(|A|,|B|,|A \cup B|)} \qquad (14)$$

, where ψ_1 is strictly increasing and ψ_2 is increasing in each of its 3 variables is a common generalization of all measures encountered here : J, D, Cos, N, O_1, O_2, R, P. In this list there are weak and strong similarity measures. In the next two theorems we will describe the properties of ψ_1 and ψ_2 that will make F a good weak or strong similarity measure. We start with the latter one.

Theorem II.1.

Let F be as in (14), with ψ_1, ψ_2 functions $\mathbb{R}^+ \to \mathbb{R}^+$ such that

(i) $\psi_1 \geq 0$, $\psi_2 \geq 0$, ψ_1 strictly increasing, ψ_2 increasing (in its 3 variables), $\psi_1(0)=0$,

(ii) $a \begin{cases} \leq x \\ < y \end{cases}$ OR $a \begin{cases} < x \\ \leq y \end{cases}$

implies $\psi_1(a) < \psi_2(x,y,z)$ for all $z \begin{cases} \geq x \\ \geq y \end{cases}$

(iii) $a \begin{cases} \leq x \\ \leq y \end{cases}$ implies $\psi_1(a) \leq \psi_2(x,y,z)$ for all $z \begin{cases} \geq x \\ \geq y \end{cases}$

(iv) $a=x=y$ implies $\psi_1(a)=\psi_2(x,y,z)$ for all $z=x=y$.

Then we have

(I) All strong similarity measures J, D, Cos, N, O_2 satisfy the above requirements

(II) F is a good strong similarity measure, i.e. satisfies the requirements (F_1), (F_2), (F_3), (F_4).

Proof :

(I) As mentioned already, all measures J, D, Cos, N, O_2 are of the type (14). We will investigate if the functions ψ_1 and ψ_2 (in each case) satisfy the properties announced in this theorem.

J Here $\psi_1(a)=a$ and $\psi_2(x,y,z)=z$, hence (i) is satisfied. Let now

$a \begin{cases} \leq x \\ < y \end{cases}$ or $a \begin{cases} < x \\ \leq y \end{cases}$. Then, since for $J : z = |A \cup B| \begin{cases} \geq |A|=x \\ \geq |B|=y \end{cases}$, we have $a < z$ and hence $\psi_1(a) < \psi_2(x,y,z)$ proving (ii). The proof of (iii) and (iv) is similar.

D Here $\psi_1(a)=a$ and $\psi_2(x,y,z)=\frac{x+y}{2}$. Hence (i) is satisfied. Let now

$a \begin{cases} \leq x \\ < y \end{cases}$ or $a \begin{cases} < x \\ \leq y \end{cases}$. This implies $a < \frac{x+y}{2}$ hence

$\psi_1(a) < \psi_2(x,y,z)$, proving (ii). The proof of (iii) and (iv) is similar.

Cos Here $\psi_1(a)=a$ and $\psi_2(x,y,z)=\sqrt{xy}$. Hence (i) is satisfied. Let now

$a \begin{cases} \leq x \\ < y \end{cases}$ or $a \begin{cases} < x \\ \leq y \end{cases}$. This implies $a < \sqrt{xy}$ proving (ii).

The proof of (iii) and (iv) is similar.

N Here $\psi_1(a)=a$ and $\psi_2(x,y,z)=\sqrt{\frac{x^2+y^2}{2}}$. Hence (i) is satisfied. Let now

$a \begin{cases} \leq x \\ < y \end{cases}$ or $a \begin{cases} < x \\ \leq y \end{cases}$. This implies $2a^2 < x^2+y^2$, hence $a < \sqrt{\frac{x^2+y^2}{2}}$

proving (ii). The proof of (iii) and (iv) is similar.

O_2 Here $\psi_1(a)=a$ and $\psi_2(x,y,z)=\max(x,y)$. Hence (i) is satisfied. Let now

$a \begin{cases} \leq x \\ < y \end{cases}$ or $a \begin{cases} < x \\ \leq y \end{cases}$. Then $a < \max(x,y)$ implying (ii). The proof

of (iii) and (iv) is similar.

(II) We now show that F satisfies (F_1), (F_2), (F_3), (F_4).

(F_1) (iii) implies, since $|A\cap B| \begin{cases} \leq |A| \\ \leq |B| \end{cases}$ that $\psi_1(|A\cap B|) \leq \psi_2(|A|,|B|,z)$

$\forall z \begin{cases} \geq |A| \\ \geq |B| \end{cases}$. Hence also for $z=|A\cup B|$. This proves $F(A,B)\leq 1$. Of course,

since ψ_1 and ψ_2 are positive, so is F.

(F_2) $F(A,B)=1$ implies

$$\psi_1(|A\cap B|)=\psi_2(|A|,|B|,|A\cup B|) \tag{15}$$

If $|A\cap B|<|A|$ or $|A\cap B|<|B|$ we see, by (ii), that

$\psi_1(|A\cap B|)<\psi_2(|A|,|B|,|A\cup B|)$, since $|A\cup B| \begin{cases} \geq |A| \\ \geq |B| \end{cases}$, contradicting (15).

Hence $|A\cap B|=|A|=|B|$

hence $A=B$.

Converseley, if $A=B$ then $|A\cap B|=|A|=|B|$, hence, by (iv)

$$\psi_1(|A\cup B|)=\psi_2(|A|,|B|,|A\cup B|)$$

since $A=B$ implies $|A\cup B|=|A|=|B|$.

Hence $F(A,B)=1$.

(F_3) $F(A,B)=0$ iff $\psi_1(|A\cap B|)=0$ iff $|A\cap B|=0$ (ψ_1 is injective and $\psi_1(0)=0$) iff $A\cap B=\emptyset$.

(F_4) This is clear from (14) and the fact that ψ_1 strictly increases. □

Theorem II.2.

Let F be as in (14), with ψ_1, ψ_2 functions $\mathbb{R}^+\to\mathbb{R}^+$ such that

(i) $\psi_1\geq 0$, $\psi_2\geq 0$, ψ_1 strictly increasing, ψ_2 increasing (in its 3 variables), $\psi_1(0)=0$ (i.e. the same condition as (i) in the previous theorem),

(ii) $a<\min(x,y,z)$ implies $\psi_1(a)<\psi_2(x,y,z)$ for all a,x,y,z

(iii) $a\leq\min(x,y,z)$ implies $\psi_1(a)\leq\psi_2(x,y,z)$ for all a,x,y,z.

Then we have

(I) All the weak similarity measures O_1, R and P satisfy the above requirements

(II) F is a good weak similarity measure, i.e. satisfies the requirements (F_1), (F'_2), (F_3), (F_4).

Proof :

(I) Again, the measures O_1, R, P are of the type (14). We will now check if the functions ψ_1, ψ_2 (in each case) satisfy the requirements in this theorem.

O_1 Here $\psi_1(a)=a$, $\psi_2(x,y,z)=\min(x,y)$. Hence (i) is satisfied.
Let now $a<\min(x,y,z)$. Hence $a<\min(x,y)$ implying $\psi_1(a)<\psi_2(x,y,z)$, hence proving (ii). The proof of (iii) is similar.

R Here $\psi_1(a)=a$, $\psi_2(x,y,z)=y$, hence (i) is satisfied.
Let now $a<\min(x,y,z)$. Hence $a<y$ and hence (ii) is satisfied. The proof of (iii) is similar.

P This proof is similar to the one of R.

(II) We will show that F satisfies (F_1), (F'_2), (F_3), (F_4).

(F_1) $F\geq 0$ since ψ_1 and ψ_2 are positive. Further, since $|A\cap B|\leq \min(|A|,|B|,|A\cup B|)$ we have, by (iii), that

$$\psi_1(|A\cap B|)\leq\psi_2(|A|,|B|,|A\cup B|)$$

hence $F(A,B)\leq 1$.

(F'_2) Let $F(A,B)=1$. Hence $\psi_1(|A\cap B|)=\psi_2(|A|,|B|,|A\cup B|)$. By (ii) and (iii) and since $|A\cap B|\leq\min(|A|,|B|,|A\cup B|)$, we have that

$$|A\cap B| = \min(|A|,|B|,|A\cup B|)$$
$$= \min(|A|,|B|)$$

implying that $A\supset B$ or $A\subset B$.

($\underline{F_3}$) $F(A,B)=0$ iff $\psi_1(|A\cap B|)=0$ iff $|A\cap B|=0$ (ψ_1 is injective and $\psi_1(0)=0$) iff $A\cap B=\emptyset$.

($\underline{F_4}$) This is clear from (14) and the fact that ψ_1 strictly increases. □

We think that these results are important since they considerably generalize the known similarity measure and also - even more importantly in the author's opinion - show the real nature of strong and weak similarity measures.

In addition to this we can deduce from it a general theory of good similarity measures (strong or weak) for fuzzy sets. Here we only have to study the single form (14), thereby generalizing all the mentioned similarity measures to fuzzy sets. This will be done in the next section.

III. Theory of weak and strong similarity measures for fuzzy sets.

Fuzzy sets are well-known in mathematics and we suffice by mentioning the elementary definitions. Readers who are interested in more facts on fuzzy sets and their applications are referred to Zadeh (1975). Ordinary sets can be described via their so-called characteristic function : for $A\subset\Omega$, $\chi_A(x)=1$ if $x\in A$ and $\chi_A(x)=0$ if $x\notin A$. Hence the range of χ_A is $\{0,1\}$. To allow for "fuzzy" membership (as e.g. is the case in an ordered set : elements with high ranks belong "less" to the set than the ones with low ranks) we now allow the membership function to range in the interval $[0,1]$.

Hence a fuzzy set is an ordinary set $A \subset \Omega$, equiped with a function

$$P_A : \Omega \to [0,1]$$

where $P_A(x)=0$, $\forall x \in \Omega \backslash A$. Note that the case of ordinary sets is contained in this model : here $P_A(x)=1 \Leftrightarrow x \in A$.

Fuzzy intersection of the fuzzy sets A and B is the ordinary intersection $A \cap B$, with the membership function.

$$P_{A \cap B}(x) = \min(P_A(x), P_B(x)). \tag{16}$$

Fuzzy union of the fuzzy sets A and B is the ordinary union $A \cup B$, with the membership function

$$P_{A \cup B}(x) = \max(P_A(x), P_B(x)). \tag{17}$$

For A and B fuzzy sets, we say that $A \subset B$ (fuzzy inclusion) if

$$P_A(x) \leq P_B(x) \tag{18}$$

$\forall x \in \Omega$ (in fact it is enough to require this $\forall x \in A$ as is easily seen). If $A \subset B$ and $B \subset A$ (fuzzy inclusion) then we say that $A=B$ (fuzzy equality).

The extension of the cardinality of a set, to fuzzy sets is as follows :

$$|A| = \sum_{x \in A} P_A(x) \tag{19}$$

Note that all these definitions boil down to the classical ones in case of ordinary sets.

In Buell and Kraft (1981a,b) these notions were used to define R and P for fuzzy sets. They also formulated the conjecture that "rank-order comparison measures might be more appropriate for evaluation". We give an answer to this conjecture in Egghe and Michel (2002) and in this paper : the rank-order approach works well for strong similarity measures (Egghe and Michel (2002)) (but not for weak similarity measures) and in this paper we will show that the fuzzy approach works for both weak and strong similarity measures (and in fact is applicable to every measure that we encountered).

In fact, defining the fuzzy variants of the weak and strong similarity measures is very easy : they are simply (14), interpreted in the fuzzy way : let A and B be fuzzy sets. We define

$$F^*(A,B) = \frac{\psi_1(|A \cap B|)}{\psi_2(|A|,|B|,|A \cup B|)} \tag{20}$$

where we use definitions (16), (17), (19) and where ψ_1 and ψ_2 are as in the previous section. Hence (20) is

$$F^*(A,B) = \frac{\psi_1\left(\sum_{x \in A \cap B} \min(P_A(x),P_B(x))\right)}{\psi_2\left(\sum_{x \in A} P_A(x), \sum_{x \in B} P_B(x), \sum_{x \in A \cup B} \max(P_A(x),P_B(x))\right)} \tag{21}$$

We say that F^* is the fuzzy similarity measure, derived from F. We will prove that F^* is a good weak or strong similarity measure under the same conditions as in theorem II.1 and II.2 for ordinary sets. The definitions of good strong and weak similarity measures for fuzzy sets are the same as in the case of ordinary sets :

Definition III.1.

We say that F^* is a strong similarity measure for fuzzy sets if, for all A,B fuzzy sets :

(F^*_1) $0 \leq F^*(A,B) \leq 1$ (22)

(F^*_2) $F^*(A,B) = 1 \Leftrightarrow A = B$ (23)

(F^*_3) $F^*(A,B) = 0 \Leftrightarrow A \cap B = \emptyset$ (24)

(F^*_4) If the denominator of F^* is constant then F^* is strictly increasing with $|A \cap B|$.

Definition III.2.

We say that F^* is a weak similarity measure for fuzzy sets if F^* satisfies (F^*_1), (F^*_3), (F^*_4) above and

(F^*_2') $F^*(A,B)=1 \Rightarrow A \subset B$ or $B \subset A$

All these notations have to be interpreted in the fuzzy way.

We have the following results.

Theorem III.3.

Let F^* be as in (20) and let ψ_1, ψ_2 have the same properties as in theorem II.1. Then F^* is a good strong similarity measure for fuzzy sets.

Proof :

(F^*_1) Obviously $F^* \geq 0$. Since $|A \cap B| \begin{cases} \leq |A| \\ \leq |B| \end{cases}$, we have that

$\psi_1(|A \cap B|) \leq \psi_2(|A|, |B|, |A \cup B|)$ since $|A \cup B| \begin{cases} \geq |A| \\ \geq |B| \end{cases}$. We use here the definitions (16)-(19). Hence $F^* \leq 1$.

(F^*_2) Let $F^*(A,B)=1$. Hence $\psi_1(|A \cap B|)=\psi_2(|A|, |B|, |A \cup B|)$

Hence

$$|A \cap B| = |A| = |B| \tag{25}$$

by (ii) and (iii) in theorem II.1. (25) means

$$\sum_{x \in A \cap B} \min(P_A(x), P_B(x)) = \sum_{x \in A} P_A(x) = \sum_{x \in B} P_B(x).$$

Since $\forall x \in A \cap B : \min(P_A(x), P_B(x)) \begin{cases} \leq P_A(x) \\ \leq P_B(x) \end{cases}$

and

$$\forall x \in A \setminus (A \cap B) : P_A(x) \geq 0$$
$$\forall x \in B \setminus (A \cap B) : P_B(x) \geq 0$$

we have that

$$\forall x \in A \cap B : \min(P_A(x), P_B(x)) = P_A(x) = P_B(x) \tag{26}$$

and

$$\forall x \in A \setminus (A \cap B) : P_A(x) = 0 \tag{27}$$
$$\forall x \in B \setminus (A \cap B) : P_B(x) = 0. \tag{28}$$

(27) and (28) imply that $A \cap B = A = B$ as ordinary sets and then (26) implies that $A = B$ as fuzzy sets.

Conversely, if $A = B$ as fuzzy sets then $|A \cap B| = |A| = |B|$ in the fuzzy sense. Hence (iv) in theorem II.1 yields

$$\psi_1(|A \cap B|) = \psi_2(|A|, |B|, |A \cup B|),$$

hence $F^*(A,B) = 1$

(F^*_3) $F^*(A,B) = 0$ iff $\psi_1(|A \cap B|) = 0$ iff $|A \cap B| = 0$ (ψ_1 injective and $\psi_1(0) = 0$) iff

$$\sum_{x \in A \cap B} \min(P_A(x), P_B(x)) = 0.$$

This is equivalent with : $\forall x \in A \cap B$:

$$\min(P_A(x), P_B(x)) = 0$$

$$\Leftrightarrow A \cap B = \emptyset$$

(F^*_4) is trivial. □

Theorem III.4.

Let F^* be as in (20) and let ψ_1, ψ_2 have the same properties as in theorem II.2. Then F^* is a good weak similarity measure for fuzzy sets.

Proof :

(F^*_1) Obviously $F^* \geq 0$. Since $|A \cap B| \leq \min(|A|, |B|, |A \cup B|)$, (iii) in theorem II.2 implies

$$\psi_1(|A \cap B|) \leq \psi_2(|A|, |B|, |A \cup B|)$$

hence $F^* \leq 1$.

(F^*_2) $F^*(A,B) = 1$ implies $\psi_1(|A \cap B|) = \psi_2(|A|, |B|, |A \cup B|)$. Properties (ii) and (iii) in theorem II.2 imply

$$\begin{aligned} |A \cap B| &= \min(|A|, |B|, |A \cup B|) \\ &= \min(|A|, |B|) \end{aligned}$$

obviously. Hence $|A \cap B| = |A|$ or $|A \cap B| = |B|$. In the first case we have

$$\sum_{x \in A \cap B} \min(P_A(x), P_B(x)) = \sum_{x \in A} P_A(x). \tag{29}$$

Since

$$\forall x \in A \cap B : \min(P_A(x), P_B(x)) \leq P_A(x)$$
$$\forall x \in A \setminus (A \cap B) : 0 \leq P_A(x)$$

it follows from (29) that

$$\forall x \in A \cap B : \min(P_A(x), P_B(x)) = P_A(x) \tag{30}$$
$$\forall x \in A \setminus (A \cap B) : P_A(x) = 0. \tag{31}$$

(31) implies that $A \cap B = A$ as an ordinary set and (30) implies then that $A \cap B = A$ as fuzzy sets. Since $P_{A \cap B} \leq P_B$ we hence have $P_A \leq P_B$ or $A \subset B$ in the fuzzy sense. In case $|A \cap B| = |B|$ we can show in a similar way that $B \subset A$ in the fuzzy sense.

$(F^*_3), (F^*_4)$ The proof is the same as in theorem III.3. □

From theorems III.3 and II.1 it now follows that J, D, Cos, N, O_2 (formulae (1), (2), (3), (4), (6)) but now interpreted in the fuzzy sense, are good strong similarity measures for fuzzy sets. From theorems III.4 and II.1 it follows that O_1, R and P (formulae (5), (7), (8)) but now interpreted in the fuzzy sense, are good weak similarity measures for fuzzy sets.

We have now established the theory of weak and strong similarity measures for general fuzzy sets. In the next section we will apply this theory to ordered sets of documents of the type of Fig. 1.

IV. Application to ordered similarity measures.

Let $C = (C_i)_{i \in N}$ be an ordered set (chain) as described in the introduction. We consider

$$U_C = \bigcup_{i=1}^{\infty} C_i \tag{32}$$

as an ordinary set, equiped with the membership function

$$P_{U_C}(x) = \varphi(i) \Leftrightarrow x \in C_i \tag{33}$$

where $\varphi(i)$ strictly decreases in i. This means that documents in C_i have a lower weight, the higher i, a logical requirement.

There is a lot of freedom to choose the concrete φ but in this paper we will use

$$\varphi(i) = \frac{1}{2^{i-1}} \tag{34}$$

a decreasing power of the rank i. We feel that it is better to use (34) than e.g. a linear decrease in i. Indeed, there should be a larger difference in comparing C_1 with C'_2 than e.g. C_{101} with C'_{102} in case $C_1 = C_{101}$ and $C'_2 = C'_{102}$ as sets. This is also linked to the sensation law of Weber-Fechner stating that the sensation is proportional to the logarithm of the stimulus (see Egghe (1994), Egghe and Rousseau (1990)). A similar approach has been followed in Egghe and Michel (2002).

Let now $C = (C_i)_{i \in \mathbb{N}}$ and $C' = (C'_j)_{j \in \mathbb{N}}$ be two ordered sets as described in the introduction. Based on (32), (34) and (20) we define

$$\begin{aligned} Q(C,C') &= F^*(U_C, U_{C'}) \\ &= \frac{\psi_1(|U_C \cap U_{C'}|)}{\psi_2(|U_C|, |U_{C'}|, |U_C \cup U_{C'}|)} \end{aligned} \tag{35}$$

Here (35) is in the fuzzy sense, of course. We need concrete expressions for $|U_C \cap U_{C'}|$, $|U_C|, |U_{C'}|$ and $|U_C \cup U_{C'}|$. This is given in the next lemma.

Lemma IV.1.

(i) $$|U_C \cap U_{C'}| = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |C_i \cap C'_j| \min\left(\frac{1}{2^{i-1}}, \frac{1}{2^{j-1}}\right) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |C_i \cap C'_j| \frac{1}{2^{\max(i,j)-1}} \tag{36}$$

(ii) $$|U_C| = \sum_{i=1}^{\infty} |C_i| \frac{1}{2^{i-1}} \tag{37}$$

(iii) $$|U_{C'}| = \sum_{j=1}^{\infty} |C_j'| \frac{1}{2^{j-1}} \tag{38}$$

(iv) $$|U_C \cup U_{C'}| = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} |C_i \cap C_j'| \frac{1}{2^{i-1}} + \sum_{i=1}^{\infty} \sum_{j=1}^{i-1} |C_i \cap C_j'| \frac{1}{2^{j-1}} +$$

$$\sum_{i=1}^{\infty} |C_i \setminus \dot{\bigcup}_{j=1}^{\infty} C_j'| \frac{1}{2^{i-1}} + \sum_{j=1}^{\infty} |C_j' \setminus \dot{\bigcup}_{i=1}^{\infty} C_i| \frac{1}{2^{j-1}} \tag{39}$$

Here all $|.|$ involving C_i, C_j' or combinations thereoff are in the ordinary sense (i.e. number of elements in) since the sets C_i, C_j' are ordinary sets, and $\dot{\bigcup}$ denotes disjoint union.

Proof :

The proofs of (i), (ii), (iii) are obvious from the definition of the fuzzy sets U_C and $U_{C'}$ (note that the ordinary sets $C_i \cap C_j'$ $(i,j \in \mathbb{N})$, C_i $(i \in \mathbb{N})$, C_j' $(j \in \mathbb{N})$ are disjoint).

The proof of (iv) requires more work. We have

$$U_C \cup U_{C'} = \left(\dot{\bigcup}_{i=1}^{\infty} C_i \right) \cup \left(\dot{\bigcup}_{j=1}^{\infty} C_j' \right) \tag{40}$$

Here $\dot{\bigcup}$ denotes the union of disjoint sets and all unions are meant in the normal (non fuzzy way). The problem is to write (40) as a disjoint union of (ordinary) sets. This way it will be easier to calculate the membership functions for $U_C \cup U_{C'}$. We have

$$\left(\dot{\bigcup}_{i=1}^{\infty} C_i \right) \cup \left(\dot{\bigcup}_{j=1}^{\infty} C_j' \right) = \dot{\bigcup}_{i=1}^{\infty} \dot{\bigcup}_{j=1}^{\infty} (C_i \cap C_j') \dot{\cup} \dot{\bigcup}_{i=1}^{\infty} \left(C_i \setminus \dot{\bigcup}_{j=1}^{\infty} C_j' \right) \dot{\cup} \dot{\bigcup}_{j=1}^{\infty} \left(C_j' \setminus \dot{\bigcup}_{i=1}^{\infty} C_i \right) \tag{41}$$

Now

$$\dot{\bigcup_{i=1}^{\infty}}\ \dot{\bigcup_{j=1}^{\infty}}(C_i \cap C_j') = \dot{\bigcup_{i=1}^{\infty}}\ \dot{\bigcup_{j=i}^{\infty}}(C_i \cap C_j') \dot{\cup} \dot{\bigcup_{i=1}^{\infty}}\ \dot{\bigcup_{j=1}^{i-1}}(C_i \cap C_j')$$

Hence (41) becomes

$$U_C \cup U_{C'} = \dot{\bigcup_{i=1}^{\infty}}\ \dot{\bigcup_{j=i}^{\infty}}\ (C_i \cap C_j') \dot{\cup} \dot{\bigcup_{i=1}^{\infty}}\ \dot{\bigcup_{j=1}^{i-1}}\ \left(C_i \cap C_j'\right) \dot{\cup} \dot{\bigcup_{i=1}^{\infty}}\ \left(C_i \setminus \dot{\bigcup_{j=1}^{\infty}} C_j'\right) \dot{\cup} \dot{\bigcup_{j=1}^{\infty}}\ (C_j' \setminus \dot{\bigcup_{i=1}^{\infty}} C_i) \quad (42)$$

(i) For the sets $C_i \cap C_j'$ in $\dot{\bigcup_{i=1}^{\infty}}\ \dot{\bigcup_{j=i}^{\infty}}(C_i \cap C_j')$

Here each element of $C_i \cap C_j'$ receives the same weight for the membership function for $U_C \cup U_{C'}$, namely

$$\max\left(\frac{1}{2^{i-1}}, \frac{1}{2^{j-1}}\right).$$

Since $j \geq i$, we find for all $x \in C_i \cap C_j'$ as in (i)

$$P_{U_C \cup U_{C'}}(x) = \frac{1}{2^{i-1}}. \quad (43)$$

(ii) For the sets $C_i \cap C_j'$ in $\dot{\bigcup_{i=1}^{\infty}}\ \dot{\bigcup_{j=1}^{i-1}}(C_i \cap C_j')$

Here each element of $C_i \cap C_j'$ receives the same weight for the membership function for $U_C \cup U_{C'}$, namely

$$\max\left(\frac{1}{2^{i-1}}, \frac{1}{2^{j-1}}\right).$$

Since $j < i$ we find for all $x \in C_i \cap C_j'$ as in (ii)

$$P_{U_C \cup U_{C'}}(x) = \frac{1}{2^{j-1}}. \quad (44)$$

(iii) For the sets $C_i \setminus \dot{\bigcup_{j=1}^{\infty}} C_j'$ in $\dot{\bigcup_{i=1}^{\infty}}(C_i \setminus \dot{\bigcup_{j=1}^{\infty}} C_j')$

Here each element of $C_i \setminus \dot{\bigcup_{j=1}^{\infty}} C_j'$ receives the same weight for the membership function for $U_C \cup U_{C'}$, namely

$$\max\left(\frac{1}{2^{i-1}}, 0\right)$$

$$= \frac{1}{2^{i-1}}$$

Hence, for all $x \in C_i \setminus \dot{\bigcup_{j=1}^{\infty}} C_j'$ as in (iii) we have

$$P_{U_C \cup U_{C'}}(x) = \frac{1}{2^{i-1}}. \tag{45}$$

(iv) For the sets $C_j' \setminus \dot{\bigcup_{i=1}^{\infty}} C_i$ in $\dot{\bigcup_{j=1}^{\infty}} (C_j' \setminus \dot{\bigcup_{i=1}^{\infty}} C_i)$

Here each element of $C_j' \setminus \dot{\bigcup_{i=1}^{\infty}} C_i$ receives the same weight for the membership function for $U_C \cup U_{C'}$, namely

$$\max\left(\frac{1}{2^{j-1}}, 0\right)$$

$$= \frac{1}{2^{j-1}}.$$

Hence, for all $x \in C_j' \setminus \dot{\bigcup_{i=1}^{\infty}} C_i$ as in (iv) we have

$$P_{U_C \cup U_{C'}}(x) = \frac{1}{2^{j-1}}. \tag{46}$$

Since in (42) all unions are disjoint, all these values add up, leading to

$$|U_C \cup U_{C'}| = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} |C_i \cap C_j'| \frac{1}{2^{i-1}} + \sum_{i=1}^{\infty} \sum_{j=1}^{i-1} |C_i \cap C_j'| \frac{1}{2^{j-1}} +$$

$$\sum_{i=1}^{\infty} |C_i \setminus \dot{\bigcup_{j=1}^{\infty}} C_j'| \frac{1}{2^{i-1}} + \sum_{j=1}^{\infty} |C_j' \setminus \dot{\bigcup_{i=1}^{\infty}} C_i| \frac{1}{2^{j-1}}. \tag{47}$$

This concludes the proof of the lemma. □

We can now present concrete (but intricate) formulae of good weak or strong similarity measures for ordered sets (considered as fuzzy sets), based on (35) and the lemma.

I. Weak measures

I.1. Recall

We have

$$R(C,C') = \frac{|U_C \cap U_{C'}|}{|U_{C'}|}$$

$$R(C,C') = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |C_i \cap C_j'| \; \frac{1}{2^{\max(i,j)-1}}}{\sum_{j=1}^{\infty} |C_j'| \frac{1}{2^{j-1}}} \tag{48}$$

I.2. Precision

$$P(C,C') = \frac{|U_C \cap U_{C'}|}{|U_C|}$$

$$P(C,C') = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |C_i \cap C_j'| \; \frac{1}{2^{\max(i,j)-1}}}{\sum_{i=1}^{\infty} |C_i| \frac{1}{2^{i-1}}} \tag{49}$$

I.3. Overlap measure O_1

$$O_1(C,C') = \frac{|U_C \cap U_{C'}|}{\min(|U_C|,|U_{C'}|)}$$

$$O_1(C,C') = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |C_i \cap C_j'| \; \frac{1}{2^{\max(i,j)-1}}}{\min\left(\sum_{i=1}^{\infty} |C_i| \frac{1}{2^{i-1}}, \sum_{j=1}^{\infty} |C_j'| \frac{1}{2^{j-1}}\right)} \tag{50}$$

II. Strong measures

II.1. Jaccard J

$$J(C,C') = \frac{|U_C \cap U_{C'}|}{|U_C \cup U_{C'}|}$$

$$J(C,C') = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\ \frac{1}{2^{\max(i,j)-1}}}{\alpha} \tag{51}$$

where

$$\alpha = \sum_{i=1}^{\infty}\sum_{j=i}^{\infty}|C_i \cap C_j'|\ \frac{1}{2^{i-1}} + \sum_{i=1}^{\infty}\sum_{j=1}^{i-1}|C_i \cap C_j'|\ \frac{1}{2^{j-1}} +$$

$$\sum_{i=1}^{\infty}|C_i \setminus \dot{\bigcup_{j=1}^{\infty}} C_j'|\ \frac{1}{2^{i-1}} + \sum_{j=1}^{\infty}|C_j' \setminus \dot{\bigcup_{i=1}^{\infty}} C_i|\ \frac{1}{2^{j-1}}$$

II.2. Dice D

$$D(C,C') = \frac{2|U_C \cap U_{C'}|}{|U_C| + |U_{C'}|}$$

$$D(C,C') = \frac{2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\ \frac{1}{2^{\max(i,j)-1}}}{\sum_{i=1}^{\infty}|C_i|\frac{1}{2^{i-1}} + \sum_{j=1}^{\infty}|C_j'|\frac{1}{2^{j-1}}} \tag{52}$$

II.3. Cosine Cos

$$Cos(C,C') = \frac{|U_C \cap U_{C'}|}{\sqrt{|U_C||U_{C'}|}}$$

$$Cos(C,C') = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\ \frac{1}{2^{\max(i,j)-1}}}{\sqrt{\left(\sum_{i=1}^{\infty}|C_i|\frac{1}{2^{i-1}}\right)\left(\sum_{j=1}^{\infty}|C_j'|\frac{1}{2^{j-1}}\right)}} \tag{53}$$

II.4. N

$$N(C,C') = \frac{\sqrt{2}|U_C \cap U_{C'}|}{\sqrt{|U_C|^2+|U_{C'}|^2}}$$

$$N(C,C') = \frac{\sqrt{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\,\frac{1}{2^{\max(i,j)-1}}}{\sqrt{\left(\sum_{i=1}^{\infty}|C_i|\frac{1}{2^{i-1}}\right)^2+\left(\sum_{j=1}^{\infty}|C_j'|\frac{1}{2^{j-1}}\right)^2}} \qquad (54)$$

II.5. Overlap measure O_2

$$O_2(C,C') = \frac{|U_C \cap U_{C'}|}{\max(|U_C|,|U_{C'}|)}$$

$$O_2(C,C') = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|C_i \cap C_j'|\,\frac{1}{2^{\max(i,j)-1}}}{\max\left(\sum_{i=1}^{\infty}|C_i|\frac{1}{2^{i-1}},\sum_{j=1}^{\infty}|C_j'|\frac{1}{2^{j-1}}\right)} \qquad (55)$$

Note :

It can readily been checked that the measures J, D, Cos and N (and only these) satisfy the following interesting property of similarity measures for ordered sets (denote by Q one of the above mentioned measures) :

Let $i_0, j_0 \in \mathbb{N}$ be fixed arbitrarily, $i_0 \neq j_0$. Let

$$C^{(i_0)} = (C_k)_{k\in\mathbb{N}}$$
$$C'^{(j_0)} = (C_\ell')_{\ell\in\mathbb{N}}$$

be ordered sets such that $C_k \cap C_\ell' = \emptyset$, $\forall k,\ell \in \mathbb{N}$ except for $k=i_0$ and $\ell=j_0$. If we let i_0 and j_0 vary (but not the sets C_{i_0} and C'_{j_0}) then $Q(C^{(i_0)}, C'^{(j_0)})$ is strictly decreasing in $j_0 > i_0$ (i_0 fixed) and in $i_0 > j_0$ (j_0 fixed). The easy proof is left as an exercise. The above property was also studied in Egghe and Michel (2002).

V. Experimentation

V.1. Presentation of the context

The context of experimentation is the same as the one presented in Michel (2000) and Egghe and Michel (2002). The IR-system Profil-Doc (Lainé-Cruzel et al. (1996)) linked with the SPIRIT[©2] search engine (Fluhr (1997)) is used with four different filtering profiles (called T0,T1,T2 and T3) and predefined queries (approximately 600). T0 corresponds to classical interrogation without any filtering, T1, T2 and T3 correspond to interrogations where three different filtering processes are effective. The choice of four profiles is made to reproduce different IR processes or IR systems contexts. T0 is considered as the system of reference. Indeed, for each query, we compute the similarity between answers given by the system submitted to profile T0 and answers given by the system submitted to profile T1, T2 or T3. Answers given for T0 (resp. T1, T2 or T3) are called A (resp. B) if sets are considered in a usual way and C (resp. C') if sets are considered in a fuzzy way.

We draw curves of similarity results of the type of Fig. 2 :

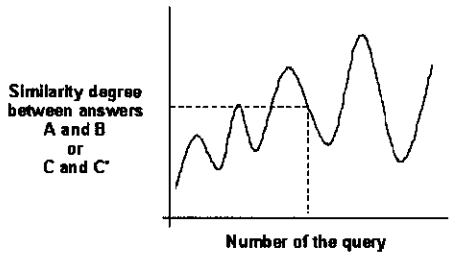

Fig. 2 A type of curve of similarity results.

Number of queries are read on the X-axis and corresponding similarity results are read on the Y-axis. The similarity is calculated according to the following measures.

---

[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 - Scientific and Technic Information Direction) - http://www.dist.cea.fr/

V.2. Measures tested

V.2.1. Fuzzy approach and classical measures

We test the five strong ordered similarity measures considered in the fuzzy context and presented before in equations (51), (52), (53), (54), (55). Measures derived from Jaccard, Dice, Cosine, N, Overlap O_2 are respectively denoted as J_F, D_F, Cos_F, N_F and O_{2F}.

We test also the three weak OS measures considered in the fuzzy context and presented in equations (48), (49), (50). Recall, Precision and Overlap O_1 are respectively denoted as R_F, P_F and O_{1F}.

Each strong fuzzy OS measure is compared with two other measures :

- its classical corresponding indicator : Jaccard, Dice, Cosine, N, Overlap O_2, respectively denoted as J, D, Cos, N, O_2 and have been presented before in equations (1), (2), (3), (4), (6).
- one strong "simple" OS measure : the one derived from Jaccard with an exponential weight (Egghe and Michel (2002)).

Each weak fuzzy OS measure is compared with :

- its classical corresponding indicator : Recall, Precision and Overlap O_1, respectively denoted as R, P, O_1 and have been presented before in equations (7), (8) and (5).
- one strong "simple" OS measure : the one derived from Jaccard with an exponential weight.

Why did we only choose this "simple" OS measure in all the comparisons ?

V.2.2. The choice of one strong "simple" OS measure

We have seen in Egghe and Michel (2002) how to construct 10 strong "simple" OS measures, derived from the 5 classical ones, by using an exponential or a linear weight. For example, the strong "simple" OS measure derived from Jaccard with an exponential weight is :

$$J_S^P(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} J(C_i,C_j')\varphi_P(i,j) \quad (56)$$

with

$$\varphi_P : \mathbb{N} \times \mathbb{N} \to \mathbb{R}^+ \; ; \; \varphi_P(i,j) = \frac{4^{m_0}}{4^{m_0} - 1} \frac{3}{2^i 2^j 2^{|i-j|}} \quad (57)$$

where $m=|C|$, $m'=|C'|$ and $m_0=\max(m,m')$.

Strong "simple" OS measures derived from Jaccard, N measure, Dice, Cosine and O_2 are called J_0^P, N_0^P, D_0^P, Cos_0^P, $O_{2_0}^P$ respectively when they have an exponential weight. Weak "simple" OS measure derived from Recall with an exponential weight is called R_0^P. Strong "simple" OS measures derived from Jaccard, N measure, Dice, Cosine and O_2 are called J_0^ℓ, N_0^ℓ, D_0^ℓ, Cos_0^ℓ, $O_{2_0}^\ell$ and R_0^ℓ respectively when they have a linear weight.

We briefly repeat some results of Egghe and Michel (2002).

We proved that

$$J_0^P = D_0^P \text{ and } J_0^\ell = D_0^\ell$$

After experimentation we have seen that :

$$J_0^P \approx Cos_0^P \approx R_0^P \text{ and } J_0^\ell \approx Cos_0^\ell \approx R_0^\ell.$$

By considering that $J \neq Cos \neq R \neq D$, one of the conclusions of the experimentation was that the weight function suppresses the particularity of classical indicators. So all exponential (resp. linear) "simple" OS measures are sensibly similar to J_0^P (resp. J_0^ℓ). Because the exponential weight φ_P (equation (57)) is of the same nature than the concrete membership function φ chosen in this paper (equation (34)), we choose J_0^P. To simplify notations, we call it J_0 in the sequel. Similarly we note Cos_0, N_0, $Dice_0$, O_{2_0}, R_0, P_0 and O_{1_0} for the "simple" OS measures J_0^P, Cos_0^P, N_0^P, $Dice_0^P$, $O_{2_0}^P$, R_0^P, P_0^P and $O_{1_0}^P$. So we compare all strong and weak measures considered in a classical and in a "fuzzy" way with J_0.

For each of the three profiles and for each of the 600 queries, we made similarity computations as explained in Fig. 2. We draw curves and group them in several ways in order to make comparisons. Firstly, we group the measure by construction in order to make appear a possible general tendency, that is, for each profile, we compare in the same figure all the classical strong measures (i.e. N, D, J, O_2, Cos) and in another one all the strong fuzzy OS measures (i.e. N_F, D_F, J_F, O_{2_F}, Cos_F). Secondly, we group measures according the native indicator considered : Jaccard's one, N measure's one, Dice's one ... in order to see if the type of construction has an influence.

V.3. Results

V.3.1. Comparison of strong measures by construction

Classical measures N, D, J, O_2, Cos are presented together in Fig. 3 for profile T1, in Fig. 5 for profile T2 and in Fig. 7 for profile T3. In order to make the curves readable, queries on the X-axis are ranked by increasing Js. Similarly, measures N_F, D_F, J_F, O_{2_F}, Cos_F are presented together in Fig. 4 for profile T1, in Fig. 6 for profile T2 and in Fig. 8 for profile T3. Queries on the X-axis are ranked by increasing J_F s. **Queries' rank in figures make only comparisons possible between curves drawed on the same figure and not those coming from different figures.**

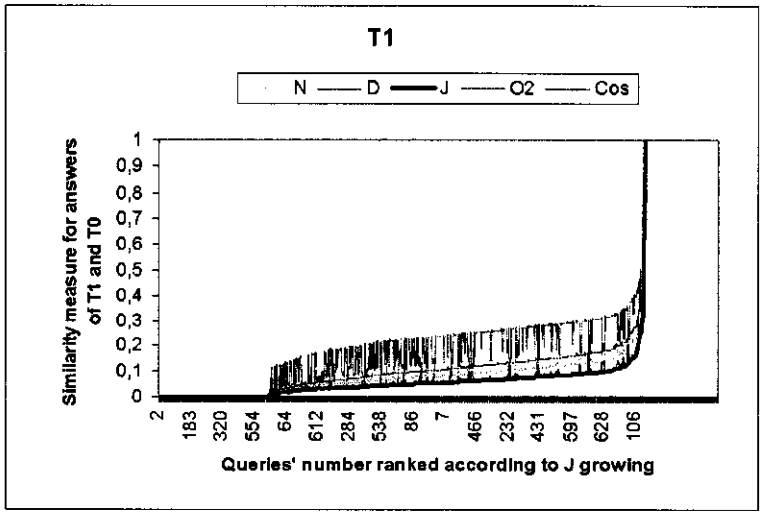

Fig. 3

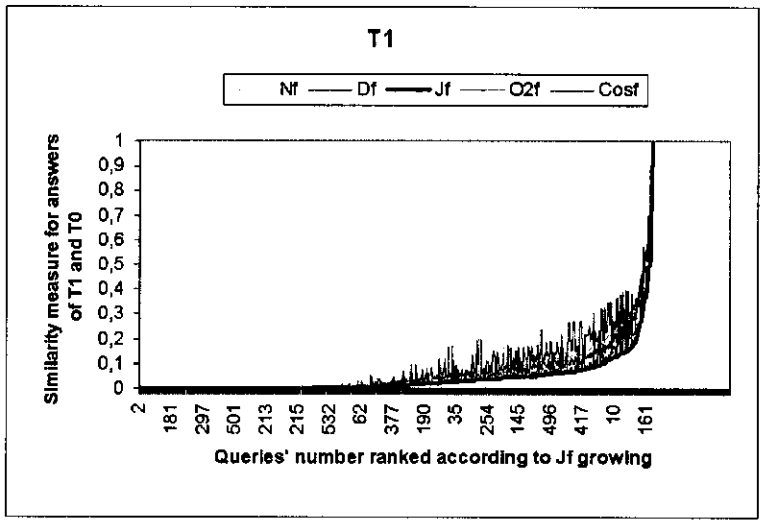

Fig. 4

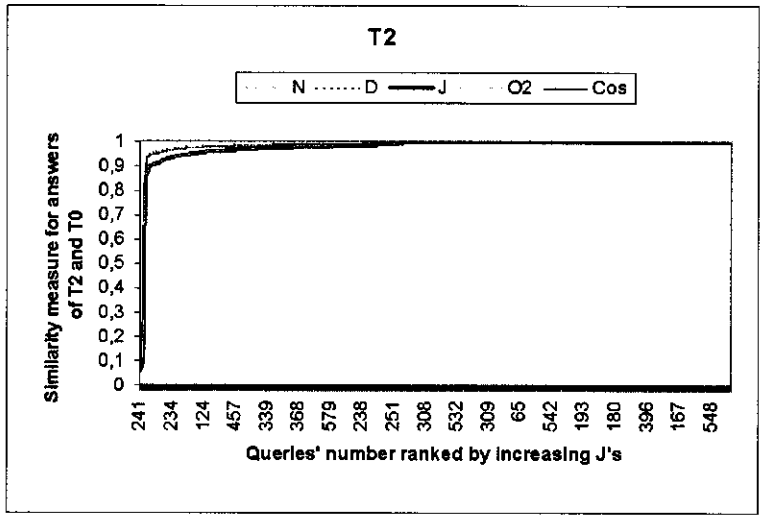

Fig. 5

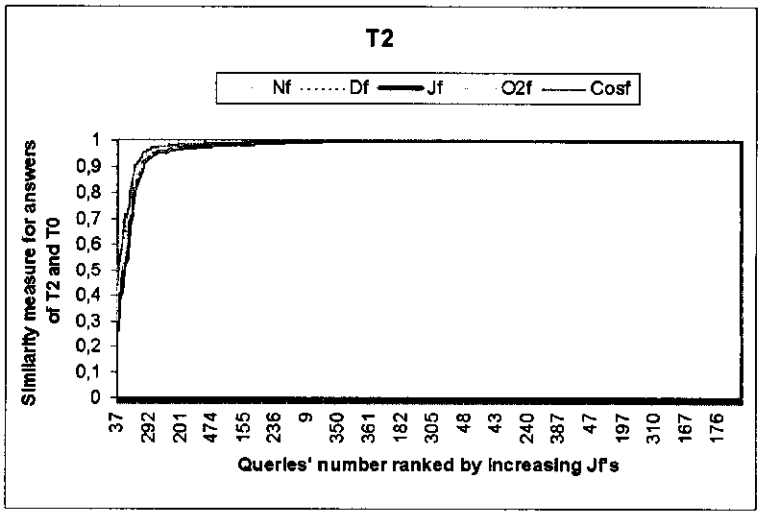

Fig. 6

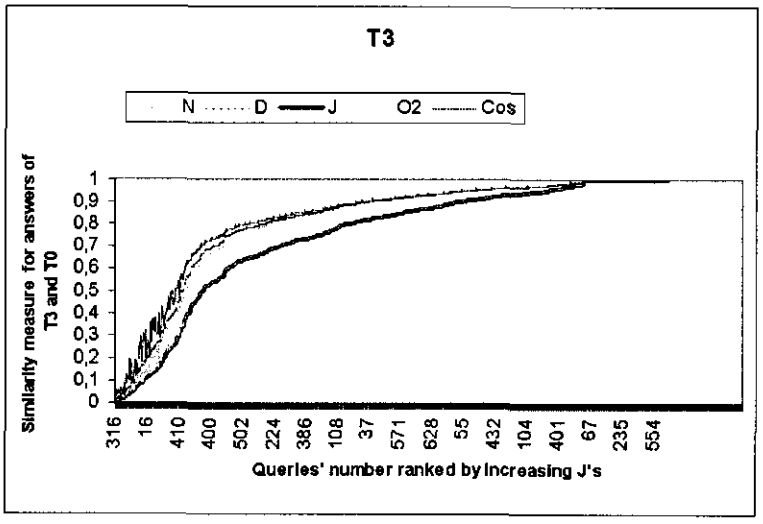

Fig. 7

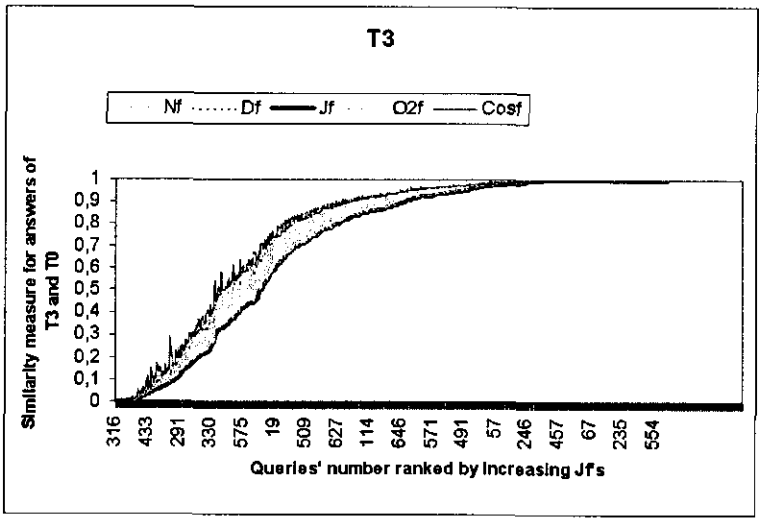

Fig. 8

We can notice that measures have specific values due to the different normalization functions ψ_2 (see Equation (14)) but share the same shape regarding the profile. Indeed, we can see a J-shape curve for profile T1 (Figs. 3, 4), an inverse J-shape curve for profile T2 (Figs. 5, 6) and an S-shape curve for profile T3 (Figs. 7, 8).

Moreover, we can see for most values that measures are ranked as $J \leq O_2 \leq D \leq N \leq Cos$ in the classical case and $J_F \leq O_{2_F} \leq D_F \leq N_F \leq Cos_F$ in the fuzzy case. Let us remember that, in the strong case (Egghe and Michel (2002)) we had $J_0 = D_0 \approx Cos_0$. **So we can say that fuzzy OS measures are more sensitive than the "simple" OS one because they respect the natural properties of classical measures i.e. the normalization function ψ_2.**

Let us compare now each strong "fuzzy" OS measure with its classical correspondent and with the strong "simple" OS measure J_0.

V.3.2. Comparison of strong measures grouped by native indicator

In the five following figures, curves of measures computed in the T1 profile case are grouped according to there native indicator. Indeed, curves of J, J_F and J_0 are presented together in Fig. 9, curves of N, N_g and J_0 (regarding experimentation results of Egghe and Michel (2002) we suppose that $N_0 \approx J_0$) are presented in Fig. 10, curves of D, D_F and J_0 (indeed $D_0 = J_0$) are presented in Fig 11, curves of O_2, O_{2_F} and J_0 (as before, se suppose that $O_{2_0} \approx J_0$) are presented in Fig. 12 and curves of Cos, Cos_F and J_0 (indeed $Cos_0 \approx J_0$) are presented in Fig. 13. In all figures, queries numbers are ranked by increasing J_F. So Figs. 3, 4 and 9 to 13 have the same rank for queries and can be compared together.

In annexes, the same curves, computed for profiles T2 (Figs. A1, A2, A3, A4 and A5) and T3 (Figs. A6, A7, A8, A9, A10), are presented.

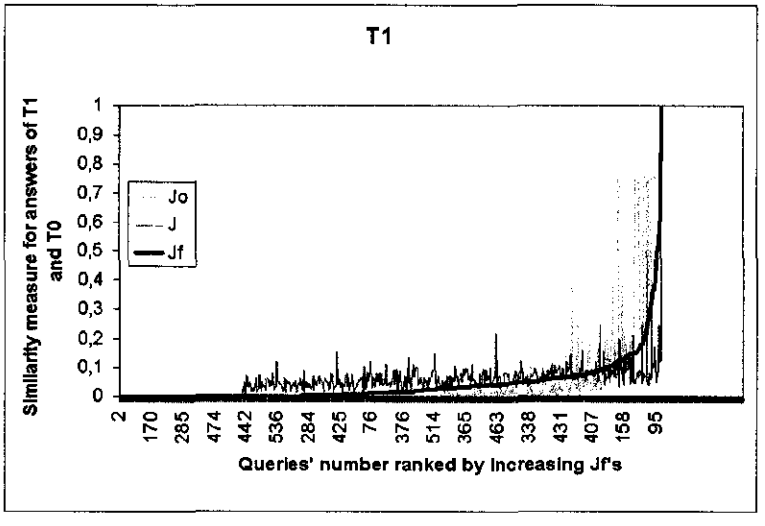

Fig. 9

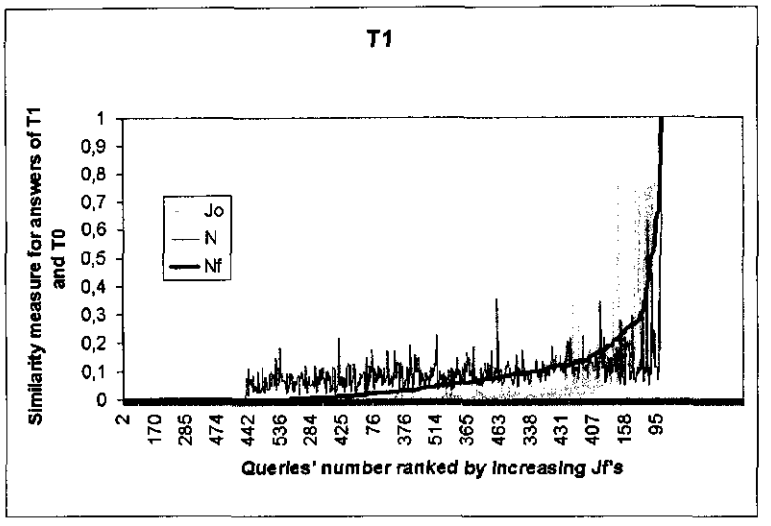

Fig. 10

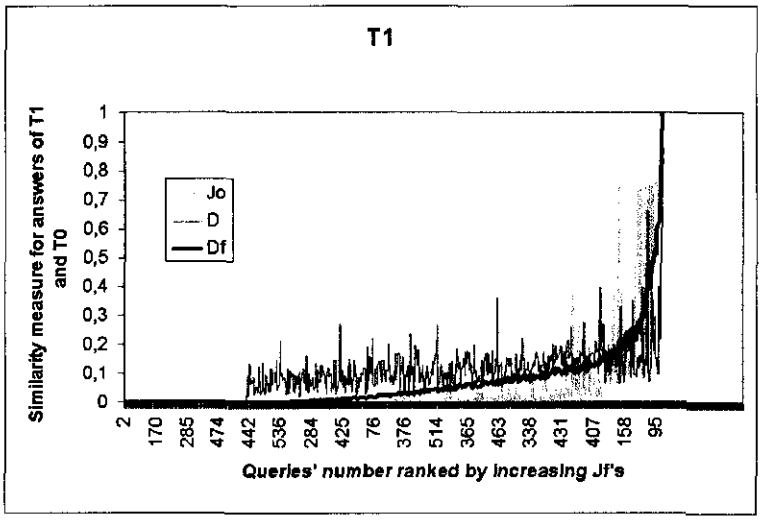

Fig. 11

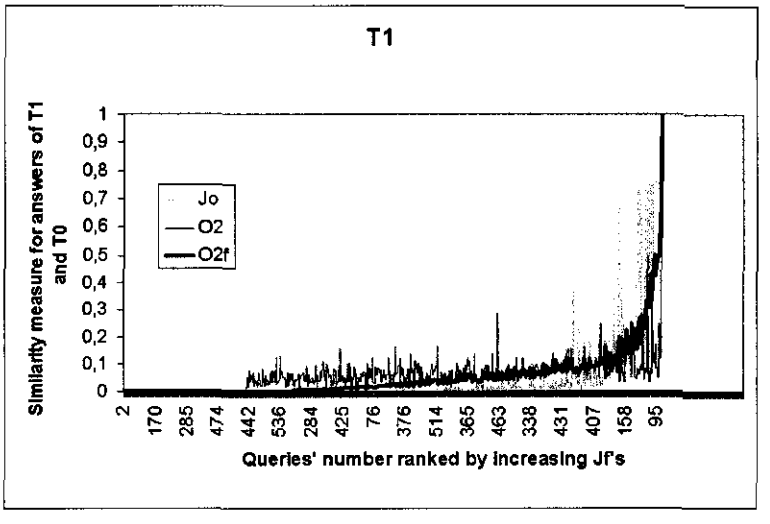

Fig. 12

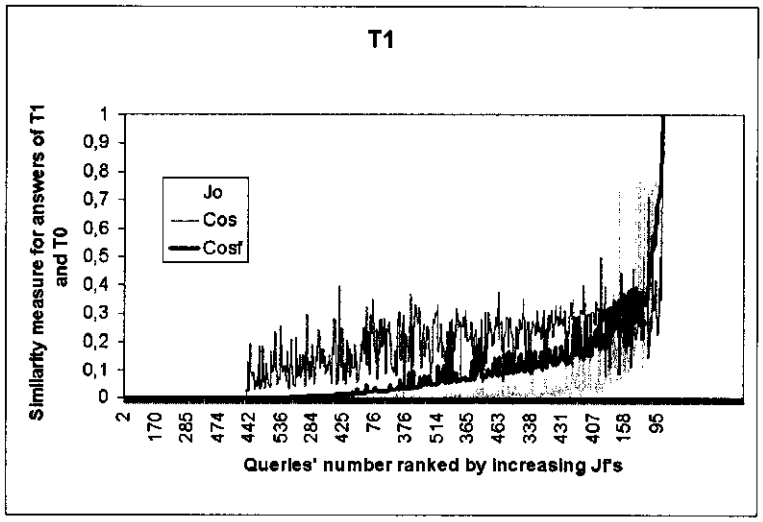

Fig. 13

The shape of the curve of Fig. 4 is naturally noticed on each figure in the "fuzzy" case (i.e. for J_F, N_F, D_F, O_{2_F} and Cos_F). On the contrary, J, N, D, O_2, Cos and J_0 do not have a regular shape. Indeed, we see that the classical curves and "simple" OS curves oscillate strongly around the fuzzy OS one. The same observations can be made by looking on figures relating to profile T2 (Figs. A1-A5) and T3 (Figs. A6-A10). So, we can say that "fuzzy" OS, "simple" OS and classical strong measures are really different.

Let us observe now the fundamental characteristics, i.e. characteristics noticed for the 3 profiles. We begin comparisons with the 3 measures J, J_F and J_0 computed for the profile T1, T2 and T3. They are presented in Figs. 9, 14 and 15 respectively. In the sequel we will see whether conclusions are extensible to other measures than the one derived from J.

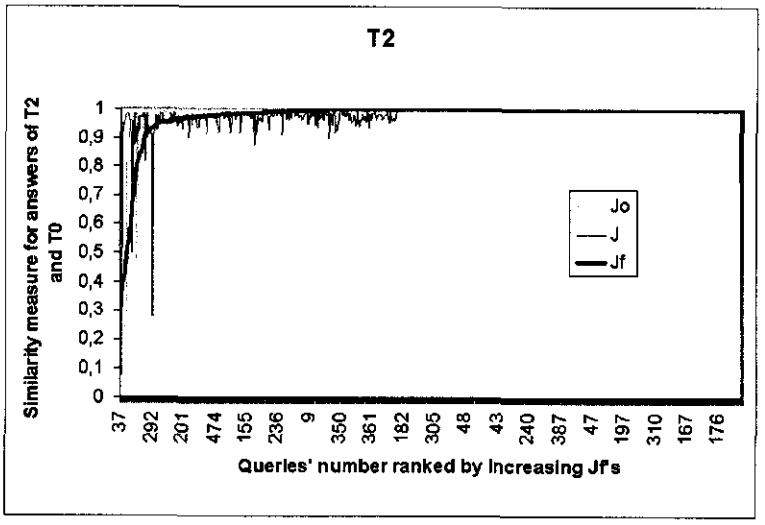

Fig. 14

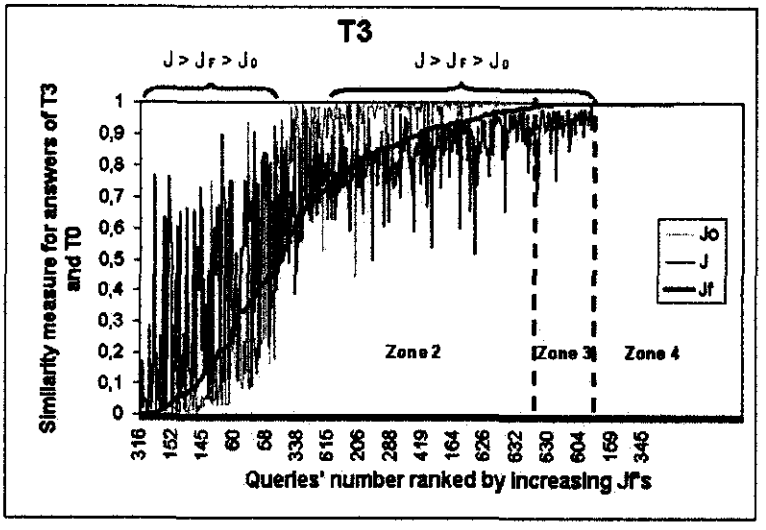

Fig. 15

The J_F curve is always very regular ; indeed queries are ranked according to increasing J_Fs. The amplitude of the oscillations of J_0 and J curves are really different depending on the profile. Indeed they are most of time very small in T1 and T2 and larger in T3. Nevertheless, we do not think that there is a characteristic of the profile but rather a characteristic depending on the J_F's distribution function. Indeed, let us define 4 zones :

- Zone 1 : J_F tends towards 0
- Zone 2 : J_F takes variable values from 0 to 1
- Zone 3 : J_F tends towards 1
- Zone 4 : J_F is equal to 1

Vertical lines in Fig. 15 make zone 2, zone 3 and zone 4 visible ; there is no zone 1 in this case. We let the reader do the same in Fig. 9 (zone 1 and 2 are noticable) and Fig. 14 (zone 2, 3 and 4 are noticable).

In zone 1 and 3, amplitude of oscillations of J_0 and J are really small and no oscillation is observed in zone 4. J_F values usually are between J_0 and J. In a more precise way in zone 1 we observe most of time that the 3 type of values are ranked as $J > J_F > J_0$ and on the contrary in zone 3 they are ranked as $J < J_F < J_0$, the permutation being effective in zone 2 (see Fig. 15). This observation is really interesting if we take into account the fact that the number of common documents is the first similarity criterion in the classical case and the rank of presentation of common documents is the one in the strong "simple" OS case. So this observation makes us believe that J_F takes into account the criteria of rank and number of common documents in a more precise and representative way than do J and J_0.

This observation is noticeable also in the figures presented in annexes. **So we can say in a general way that fuzzy OS measures are more precise and representative than classical and simple OS ones on criteria of rank and number of common documents.**

Let us observe now curves of weak measures : Recall, Precision and O_1 measures.

V.3.3. General comparison of weak measures by construction

In the six following figures we can see curves of R, P and O_1 computed for profile T1 in Fig. 16, for profile T2 in Fig. 17 and for profile T3 in Fig. 18. Curves of R_F, P_F and O_{1_F} are presented for profile T1 in Fig. 19, for profile T2 in Fig. 21 and for profile T3 in Fig. 22.

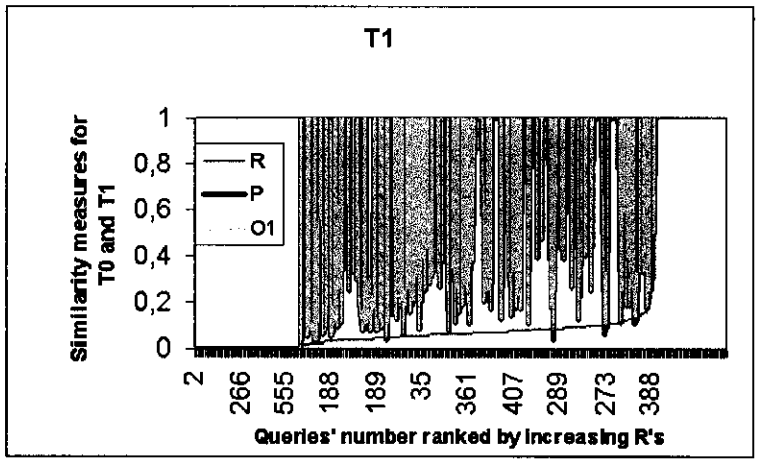

Fig. 16

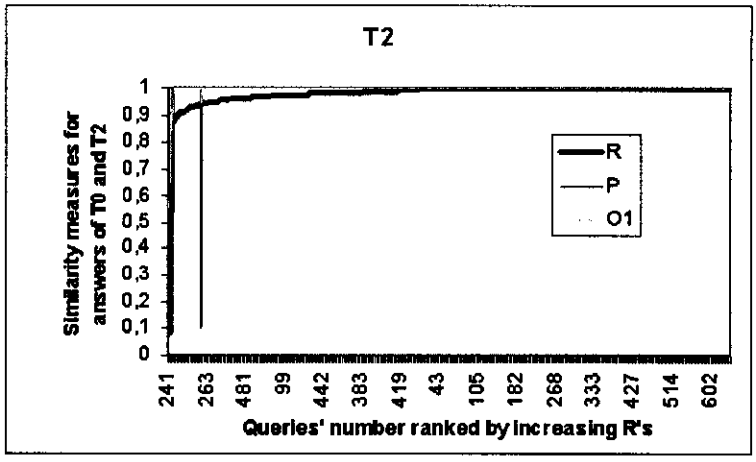

Fig. 17

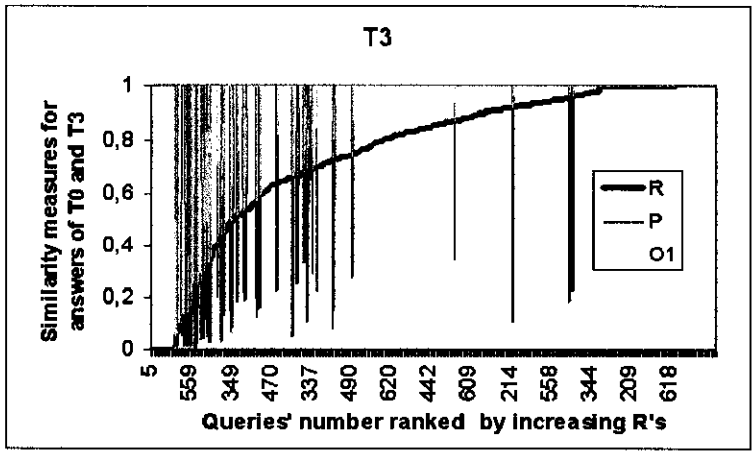

Fig. 18

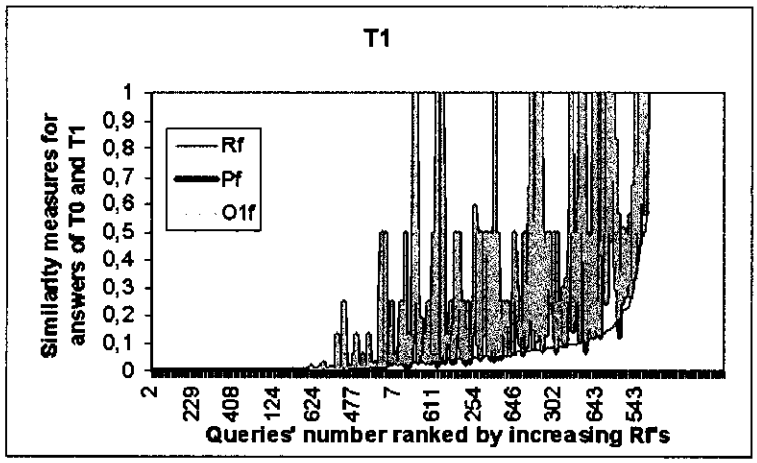

Fig. 19

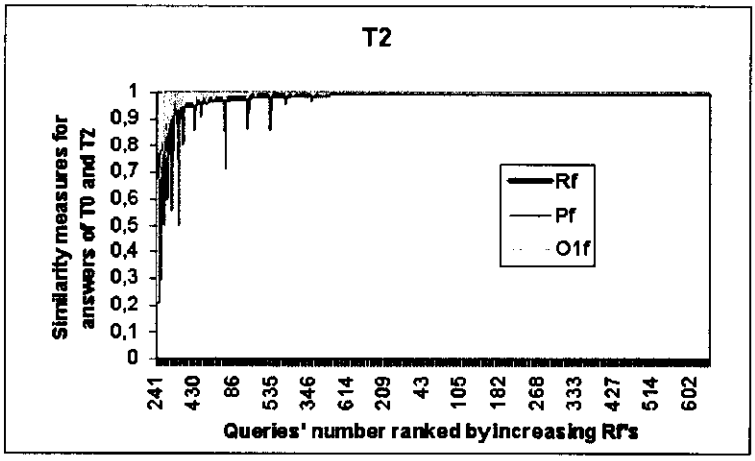

Fig. 20

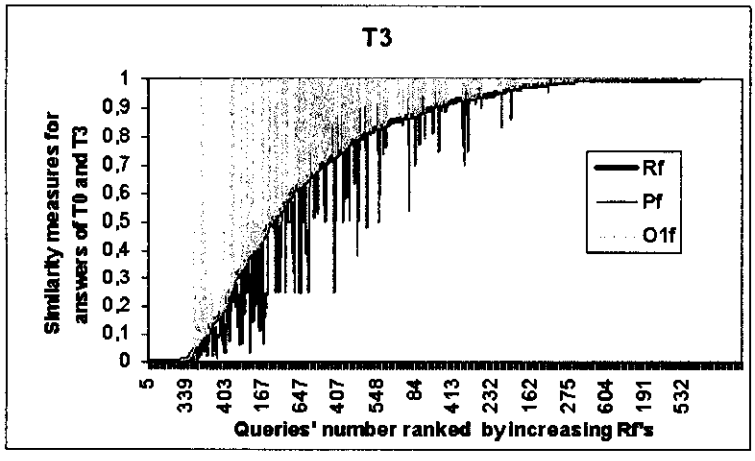

Fig. 21

We cannot notice any regularity linked with profile or measures. P and O_1 oscillate wildly around R. P_F and O_{1_F} do the same around R_F.

Previous studies have shown that, often, recall and precision vary in an inverse proportional way (see Fig. 22 or Fig. 23). It is not true in our case. Indeed we give curves (recall, precision) for each profile in annexes (Figs. A11-A16) and none of them is like Fig. 22 or Fig. 23.

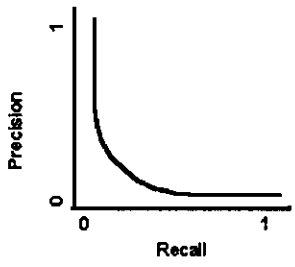

Fig. 22

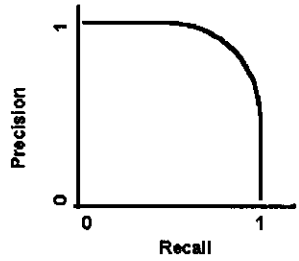

Fig. 23

This phenomenon is explained by the context of the experimentation. Indeed, recall and precision are usually computed in a collection test context. It is not the case in our experimentation where precision is calculated by taking the answers without filtering (i.e. the answer given by profile T0) as set of "retrieved documents" (corresponding to A in Equation (8) and to C in Equation (49)) and recall is calculated by taking the answers with filtering as set of "relevant documents" (corresponding to B in Equation (7) and to C' in Equation (48)). The fact that C' and B are not real sets of relevant documents in a semantic sense (regarding for example experts' judgement) is a source of biases and explain partially why we do not refind the shape of curves shown in Fig. 22 or Fig. 23. We have no conclusion by comparing R with P and R_F with P_F.

Moreover we can notice that P is very similar to O_1 and P_F is very similar to O_{1_F}. This particularity cannot be seen as a characteristic of P and O_1 because of the biases induced by data of experimentation. Indeed, if we place ourselves in the fuzzy case, C corresponds to answers linked with the neutral profile T0 and C' to answers obtained by the filtering process of profiles T1, T2 or T3. So in most of the cases we have $|C| > |C'|$, which means

that $|C'| = \min(|C|, |C'|)$ and so $P_F = O_{1_F}$. Because of the experimentation biases, we are not able to give any conclusions about weak fuzzy OS measures construction.

Nevertheless, comparisons of R with R_F and R_0 ; P with P_F and R_0, and finally O_1 with O_{1_F} and R_0 are possible by observing natural characteristics of measures. (The choice R_0 in the P or O_1 case is justified by conclusion of Egghe and Michel (2002). Indeed, we suppose that $P_0 \approx R_0$ and $O_{1_0} \approx R_0$.

V.3.4. Comparison of weak measures grouped by native indicator

Curves of measures R and R_F and R_0 are presented in Figs. 24, 25, 26 (for the profiles T1, T2 and T3 respectively). Queries are presented by increasing R_F s. Curves of measures P and P_F are presented in Figs. 27, 28, 29 (for the profiles T1, T2 and T3 respectively). Queries are presented by increasing P_F s. Curves of measures O_1 and O_{1_g} are presented in Figs. 30, 31, 32 (for the profiles T1, T2 and T3 respectively). Queries are presented by increasing O_{1_F} s.

In general, curves of R, R_0 and R_F have the same characteristics as the strong measures presented before. Zones 1, 2, 3, 4, noticed in Figs. 9, 14, 15 are present in Figs. 24, 25, 26.

On the contrary, measures P and O_1 have not the characteristics of strong measures. Indeed, oscillations of P and R_0 around P_F are larger and more irregular than the ones of R and R_0 around R_F. Observation is the same with measure O_1.

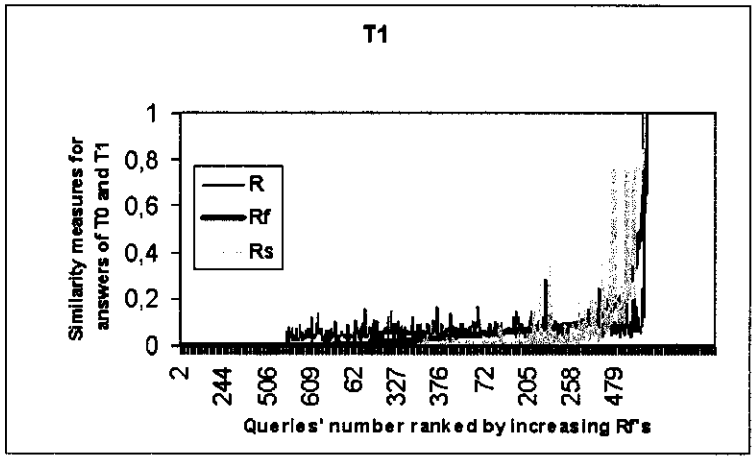

Fig. 24

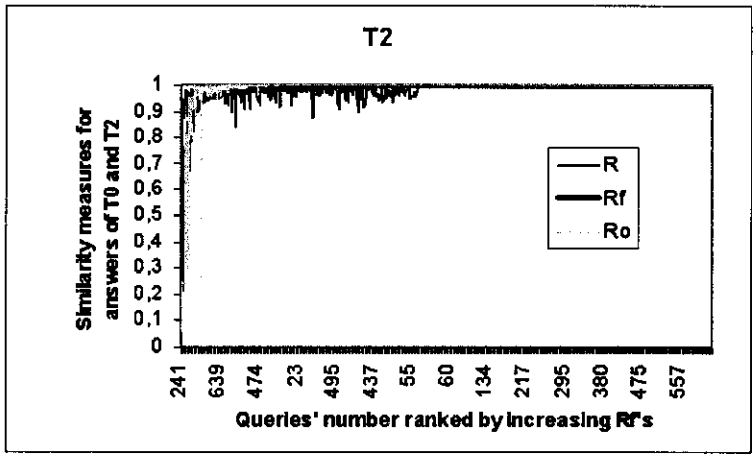

Fig. 25

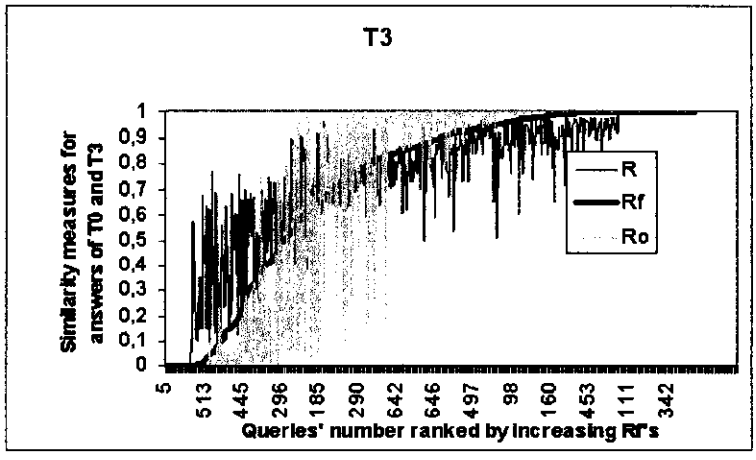

Fig. 26

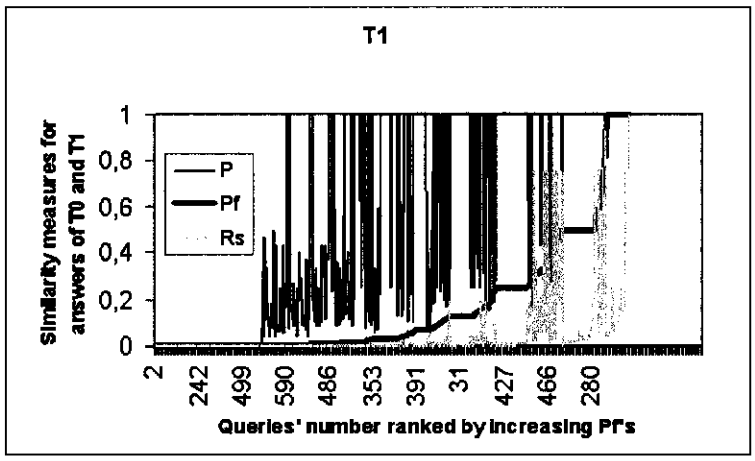

Fig. 27

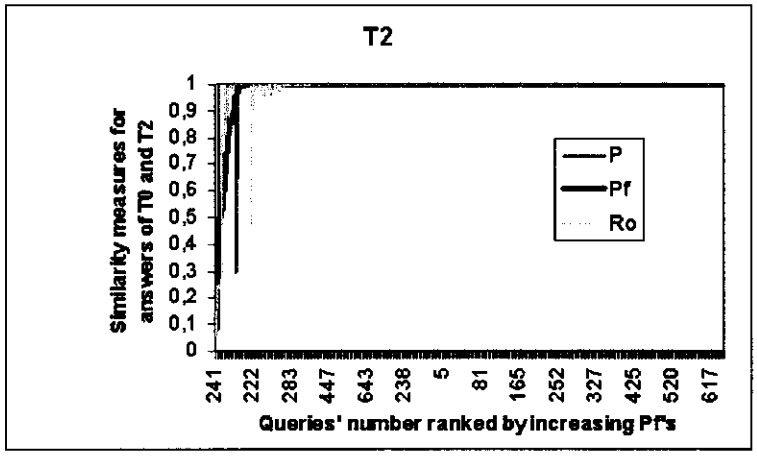

Fig. 28

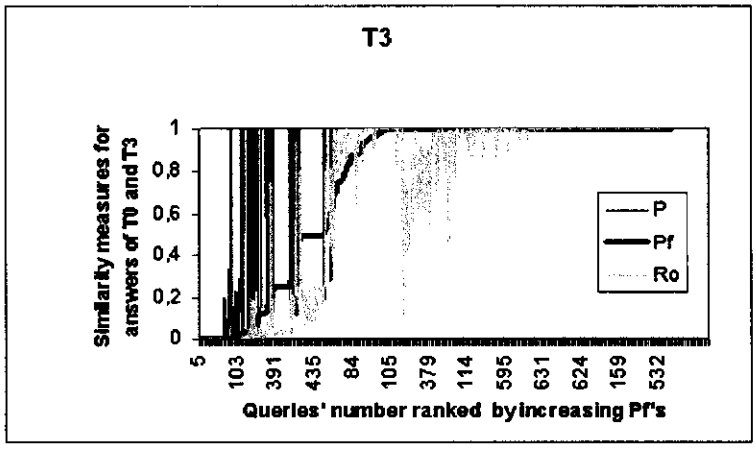

Fig. 29

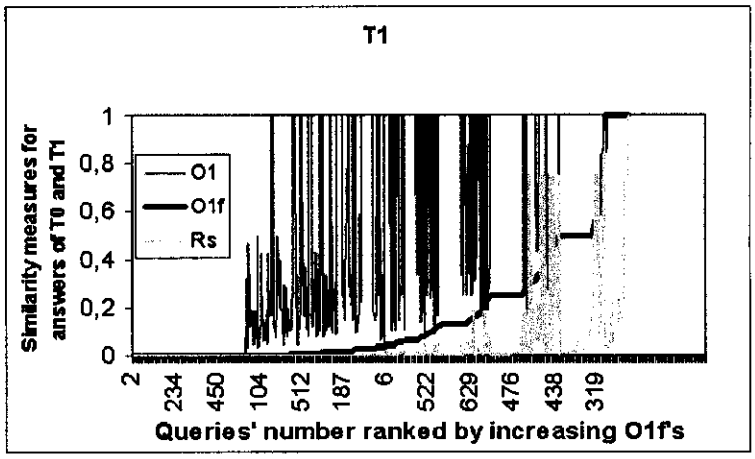

Fig. 30

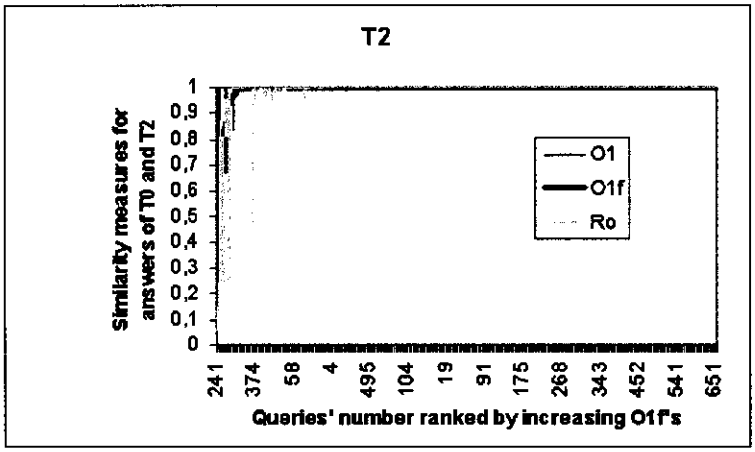

Fig. 31

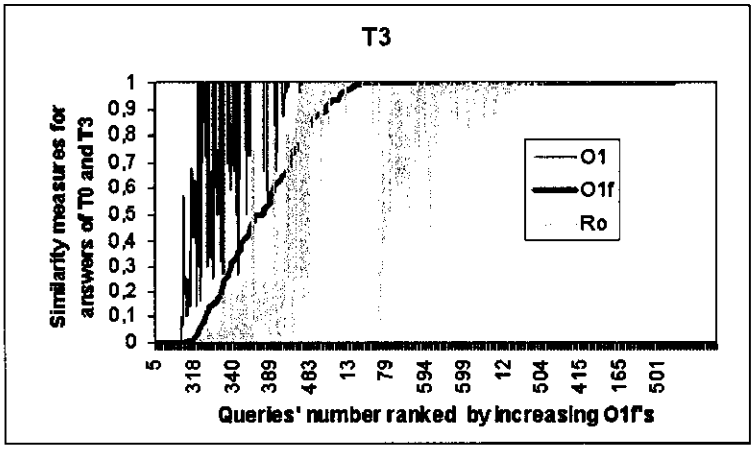

Fig. 32

V.4. Conclusion

Experimentation analyzes the characteristics of fuzzy OS measures by comparing them to classical measures and "simple" OS one. Comparisons between all the classical and all the fuzzy OS measures (i.e. comparisons by construction) are made to see general tendencies. We observe that, in the strong measures case, general shape of curves are conserved varying the profile, so the fuzzy process is stable and representative of the profile or of the IR-system used. More precisely, if we look on the values, we notice that the natural properties of classical strong measures, defined by the normalization function ψ_2, are lost in the "simple" OS case (conclusion of Egghe and Michel (2002)) but conserved in the fuzzy case. Indeed, $J \leq O_2 \leq D \leq N \leq Cos$ in the classical case, $J_F \leq O_{2_F} \leq D_F \leq N_F \leq Cos_F$ in the fuzzy case, but $J_0 = D_0 \approx Cos_0$ in the "simple" OS case. So the strong "fuzzy" OS measures are more sensitive than the strong "simple" OS one.

Comparisons of classical and weak fuzzy OS measures are not valid for two different reasons. Firstly, in the analysis of recall and precision, we attempted to find a usual regularity but we did not observe any for R and P or for R_F and P_F. We suppose that this "anomaly" is a bias of the experimentation. Indeed, recall and precision regularities are observed for collection test evaluations, which is not our context. Secondly, in the analysis of precision and O_1, we have showed that, for most values, $P_F = O_{1_F}$ and $P = O_1$, which is not normal and is due to the experimentation context. So, to give pertinent conclusions on the general tendency of weak fuzzy OS measures we need to work on a real collection test evaluation context.

In the second step, we group measures according the native indicator considered, in order to see if the type of construction has an influence on the measure or not. Theses comparisons show that, for the five strong measures and the three weak ones, fuzzy OS values are most of the time between "simple" OS values and classical values. Considering that "simple" OS measures are principally depending on the rank of common documents and classical measures are only depending on the number of common documents, we can believe that, strong (resp. weak) fuzzy OS measures are more precise and representative on

the two criteria of rank and number of common documents than classical and "simple" OS strong (resp. weak) measures.

Finally, we can say that strong fuzzy OS measures seem to be better than classical and "simple" OS strong ones : they are stable, more precise and more representative. In the weak case, we must be more prudent. We have the conviction that the preceding conclusion can be extended but we cannot confirm this before having made complementary tests in a real collection test evaluation context.

References

B.R. Boyce, C.T. Meadow and D.H. Kraft (1995). Measurement in Information Science. Academic Press, New York.

D.A. Buell and D.H. Kraft (1981a). Evaluation of fuzzy retrieval systems. Proceedings of the American Society for Information Science, 298-300.

D.A. Buell and D.H. Kraft (1981b). Performance measurement in a fuzzy retrieval environment. ACM SIGIR Forum 16(1), 56-61 (In : Proceedings of the fourth international Conference on Information Storage and Retrieval, Oakland, California, May 31-June 2, 1981).

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

L. Egghe and C. Michel (2002). Strong similarity measures for ordered sets of documents in information retrieval. Information Processing and Management, to appear.

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

C. Fluhr (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.

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

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

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

C. Michel (2000). Ordered similarity measures taking into account the rank of documents. Information Processing and Management, 37(4), 603-622.

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

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

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

L. Zadeh (1975). Fuzzy sets and their Applications to cognitive and Decision Processes. Academic Press, New York.

Annexes

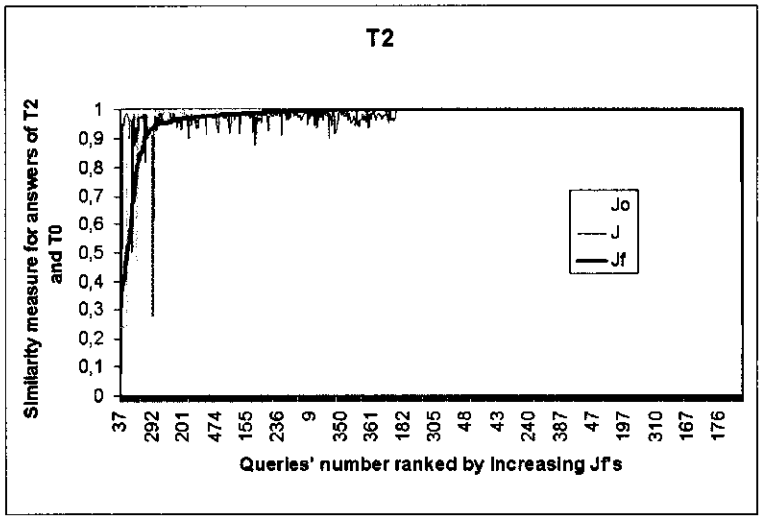

Fig. A 1

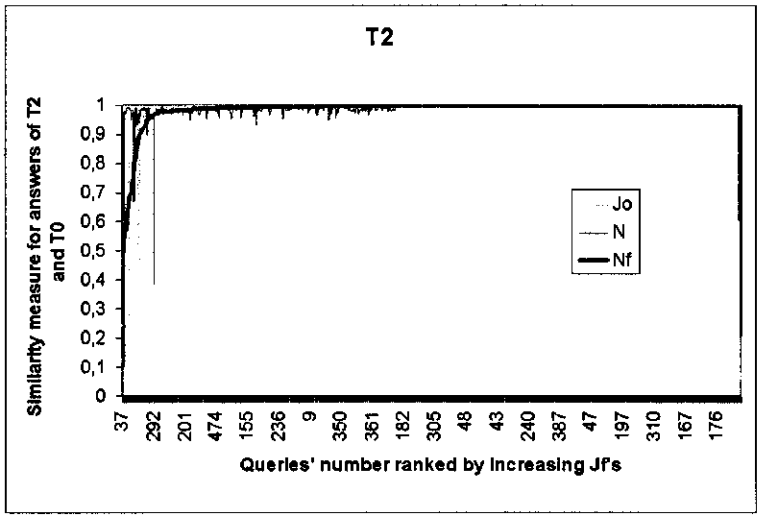

Fig. A 2

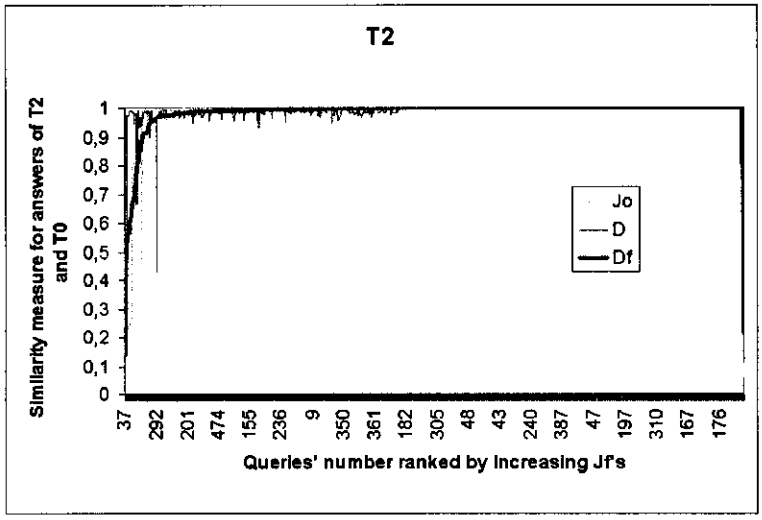

Fig. A 3

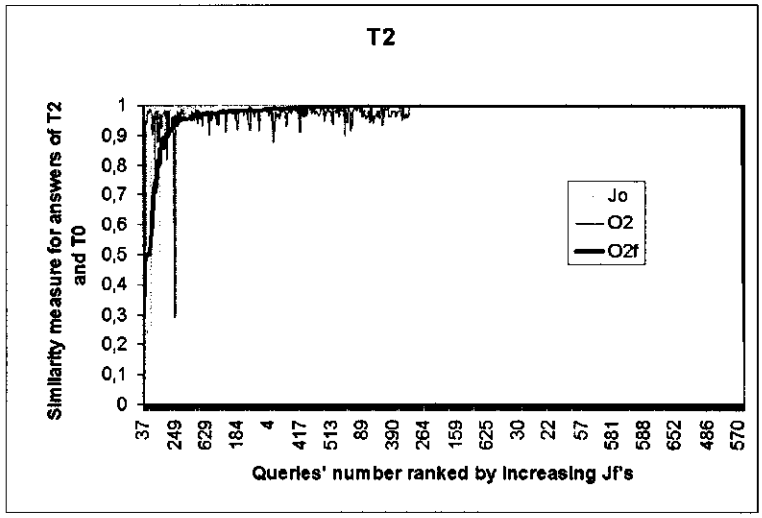

Fig. A 4

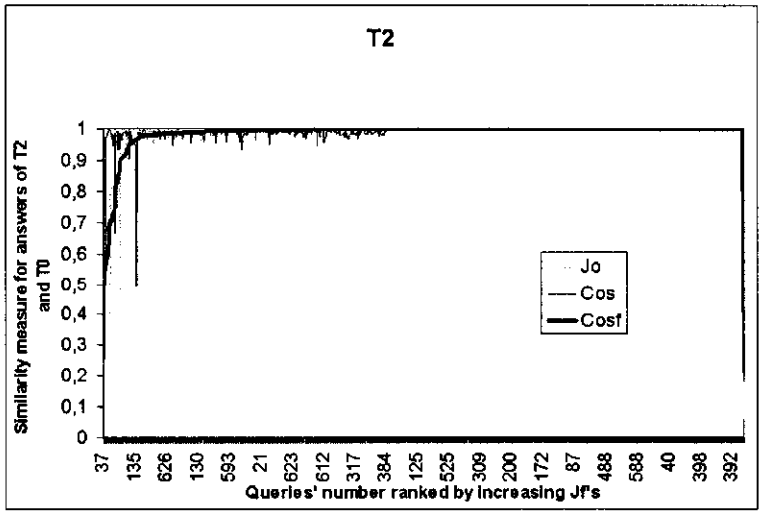

Fig. A 5

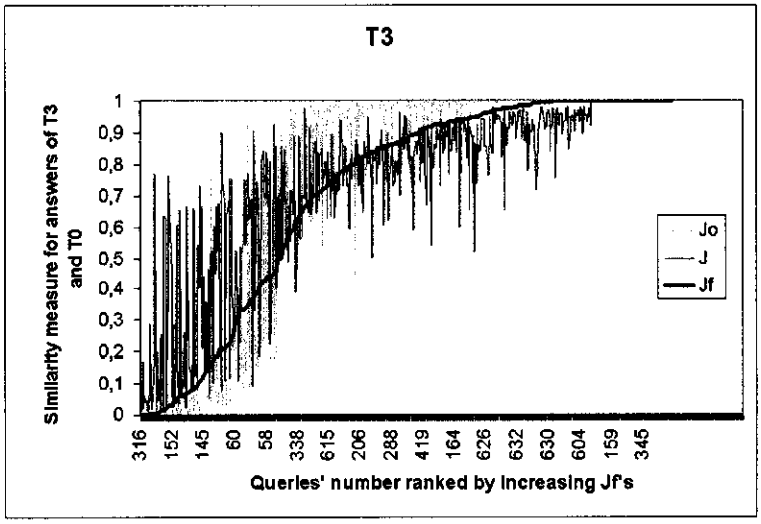

Fig. A 6

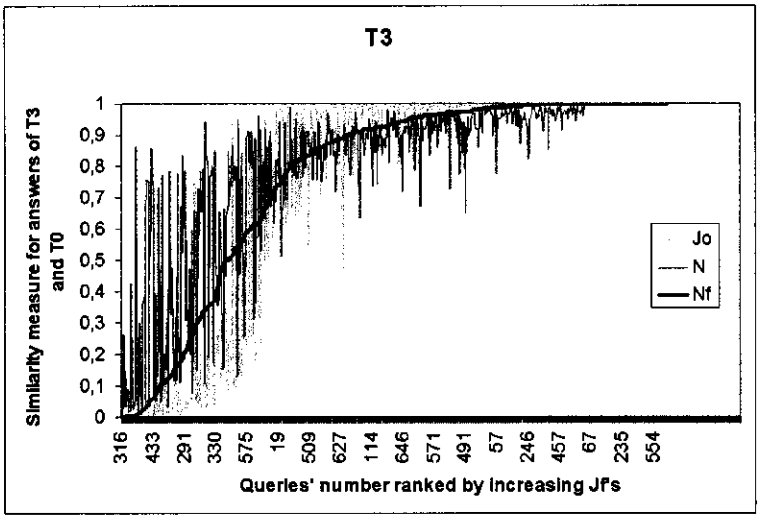

Fig. A 7

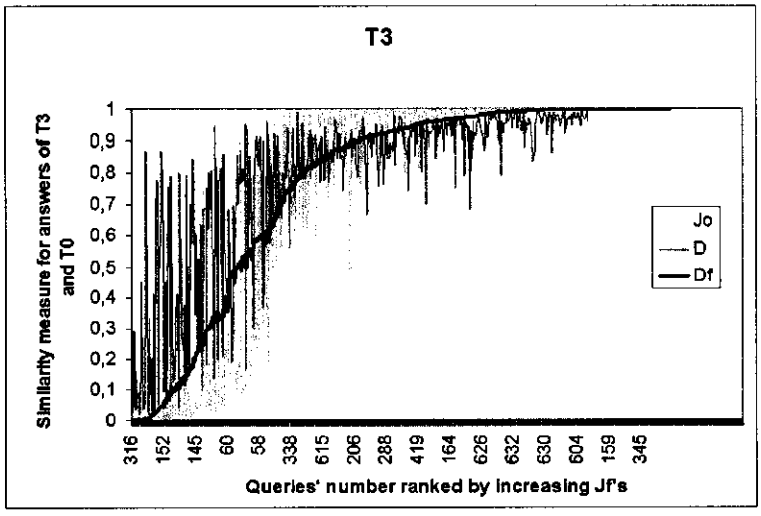

Fig. A 8

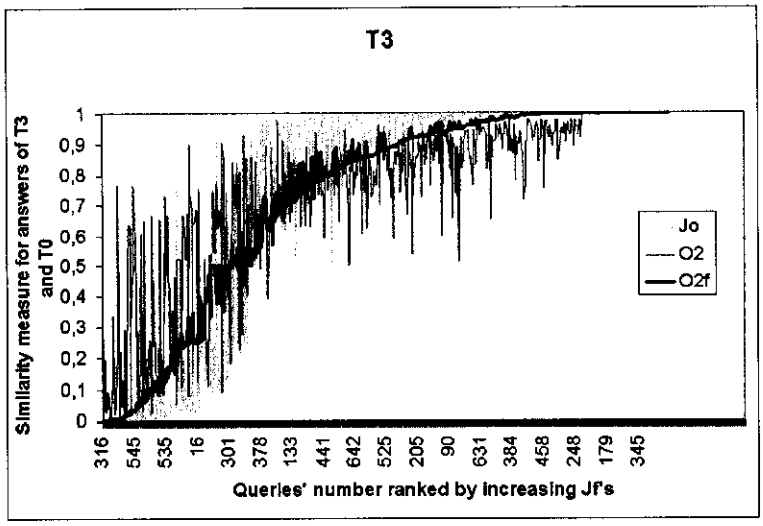

Fig. A 9

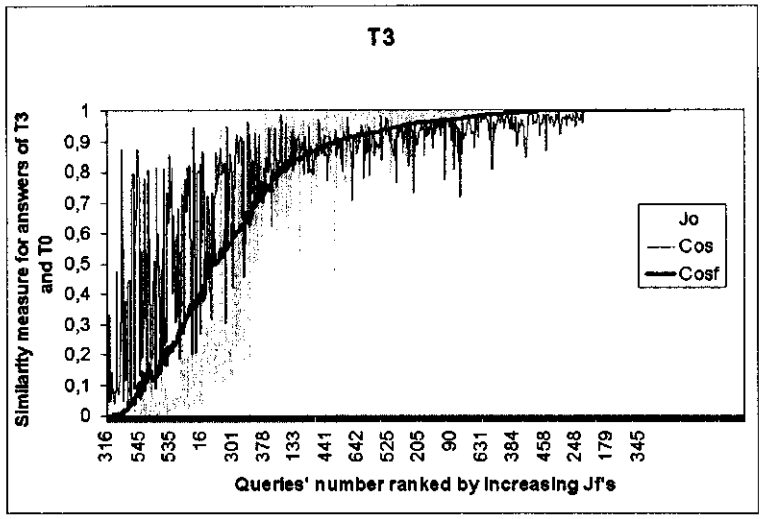

Fig. A 10

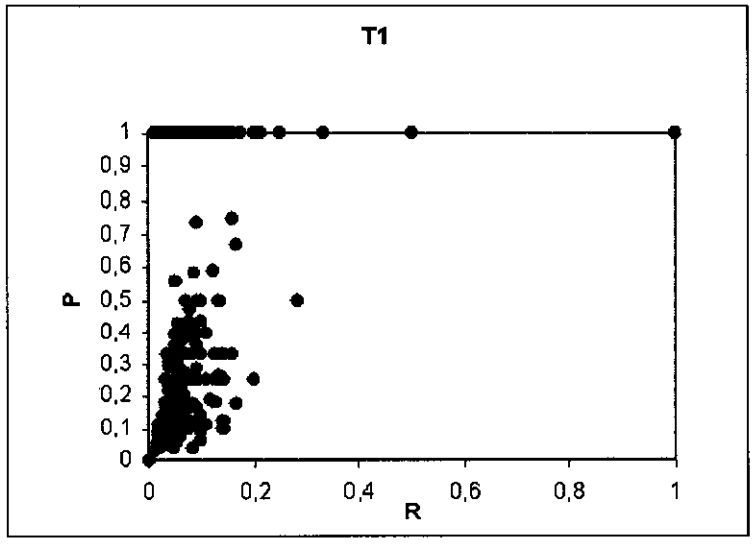

Fig. A 11

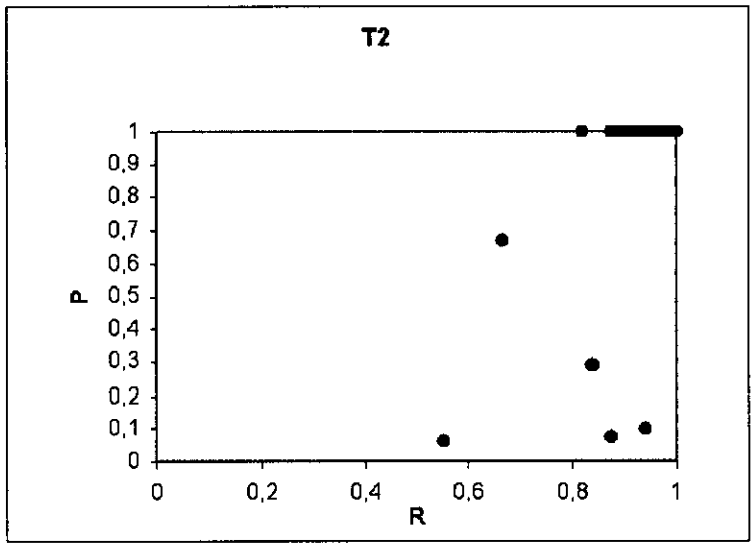

Fig. A 12

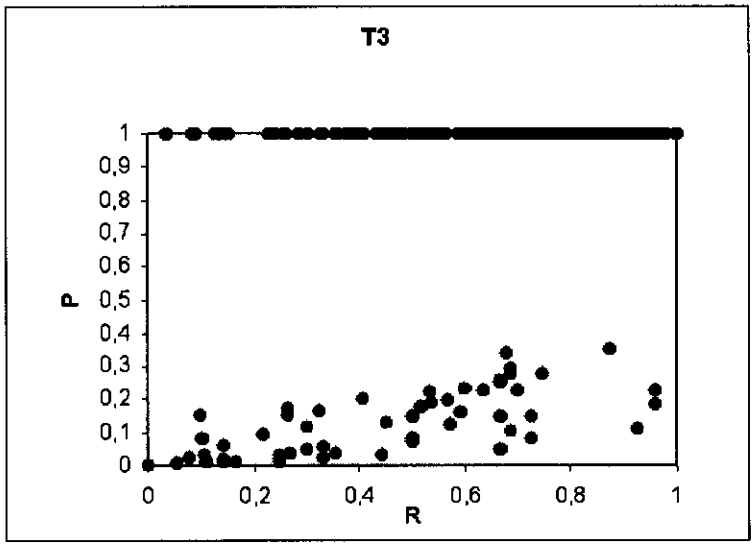

Fig. A 13

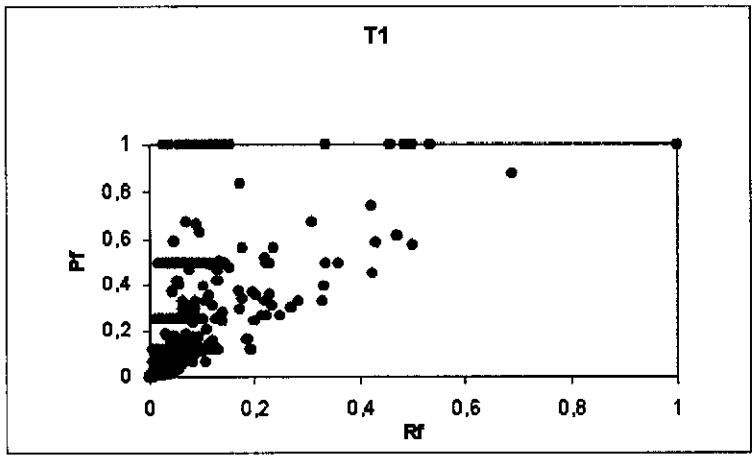

Fig. A 14

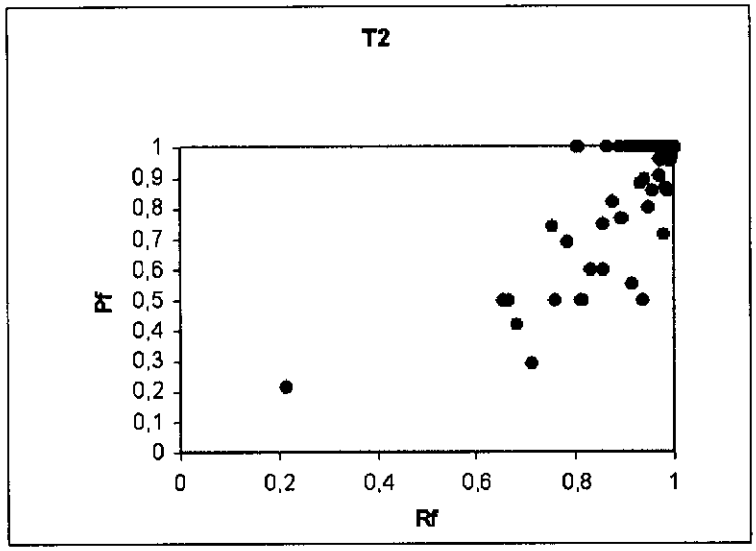

Fig. A 15

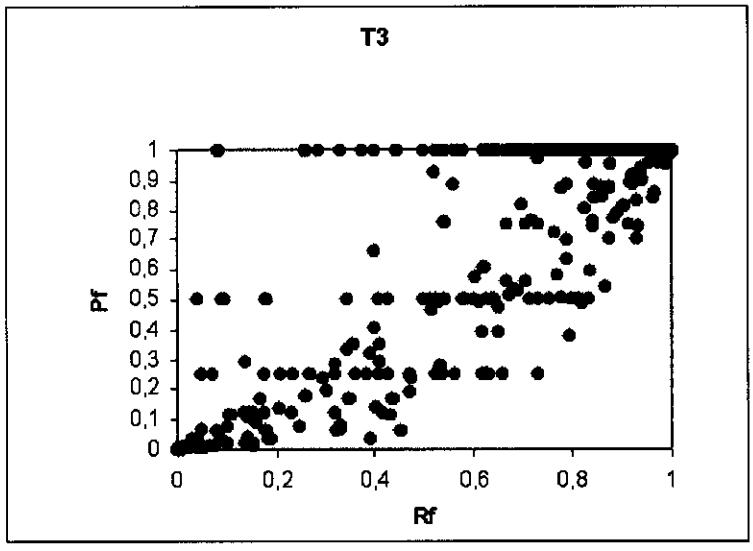

Fig. A 16