A universal method of information retrieval evaluation: the "missing" link M and the universal IR surface

EGGHE, Leo (2004) A universal method of information retrieval evaluation: the "missing" link M and the universal IR surface. In: Information Processing & Management, 40(1). p. 21-30.

DOI: 10.1016/S0306-4573(02)00094-8
Handle: http://hdl.handle.net/1942/763

A UNIVERSAL METHOD OF INFORMATION RETRIEVAL EVALUATION : THE "MISSING" LINK M AND THE UNIVERSAL IR SURFACE

by

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

ABSTRACT

The paper shows that the present evaluation methods in information retrieval (basically recall R and precision P and in some cases fallout F) lack universal comparability in the sense that their values depend on the generality of the IR problem. A solution is given by using all "parts" of the database, including the non-relevant documents and also the not-retrieved documents. It turns out that the solution is given by introducing the measure M being the fraction of the not-retrieved documents that are relevant (hence the "miss" measure). We prove that - independent of the IR problem or of the IR action - the quadruple (P,R,F,M) belongs to a universal IR surface, being the same for all IR-activities. This universality is then exploited by defining a new measure for evaluation in IR allowing for unbiased comparisons of all IR results. We also show that only using 1, 2 or even 3 measures from the set {P,R,F,M} necessary leads to evaluation measures that are non-universal and hence not capable of comparing different IR situations.

Key words : Universal IR surface, miss measure, precision, recall, fallout, silence, evaluation.

Acknowledgement : The author is grateful to H. Vanstappen for pointing out the existence of the term "silence".

[1] Permanent address.

I. Introduction

Quantitative evaluation in information retrieval is performed - very classically - by means of the measures precision (P) and recall (R), defined as follows. Let Ω denote the database (universe) in which an IR search (based on a given problem) is executed. As usual we suppose that we can determine in a unique, dichotomous way whether a document is relevant for the problem (transformed into a query according to the IR system's rules) or not. Let us denote by rel the set of relevant documents and by ret the set of retrieved documents. Then we have the following definitions :

$$P = \frac{|rel \bigcap ret|}{|ret|} \tag{1}$$

$$R = \frac{|rel \bigcap ret|}{|rel|} , \tag{2}$$

where $|.|$ denotes the cardinality of the set, i.e. the number of documents in it. Denote briefly $\neg ret = \Omega \backslash ret$ and $\neg rel = \Omega \backslash rel$, the set of not-retrieved documents and the set of non-relevant documents respectively. Sometimes (see below for references) also the following measure is considered, called fallout (F) :

$$F = \frac{|\neg rel \bigcap ret|}{|\neg rel|} \tag{3}$$

The following formula is easily derived from (1), (2) and (3) and appears e.g. in Salton and Mc Gill (1987) (p. 175) and Van Rijsbergen (1979) (p. 149)

$$P = \frac{RG}{RG+F(1-G)} , \tag{4}$$

where G denotes the "generality" of the problem :

$$G = \frac{|rel|}{|\Omega|} , \tag{5}$$

the fraction of relevant documents (describing indeed the degree of generality of the problem).

A simple, but important remark of S. Dominich (see Dominich (2001)), which is not more than a reformulation of (4), was the starting point for the results of the present paper. It goes as follows : from (4) and (5) it is easy to see that

$$\frac{FP}{R(1-P)} = \frac{|rel|}{|\Omega|-|rel|} = \frac{G}{1-G} \quad . \tag{6}$$

Formula (6) is interpreted in Dominich (2001), p. 226-227 as the effectiveness surface of IR, corresponding to a given problem or query q. He rightly points out that the shape of this surface is the same for every IR system. Keeping the problem fixed (hence fixed G), different IR techniques (e.g. relevance feedback but this is only an example) can give different values of P, R and F but they are constraint to the surface (6).

The starting observation of this paper is that, as long as we are on the same effectiveness surface, we can perfectly compare, say, two sets of IR results : (P_1, R_1, F_1) and (P_2, R_2, F_2). In other words, within the limitation of a given G (a given problem) we can compare the above triplets. Rephrased in a "negative" way, comparing two triplets as above is not possible if G varies, e.g. for different problems, since we are on different effectiveness surfaces.

Based on the above observations we can, formally, define the concepts of "surface", "semi-universal surface" and "universal surface" as follows. The concept of surface (in k-dimensional space $\mathbb{R}^k$, here k=4) is well-known and is described by an equation of the form

$$f(x,y,z) = C \tag{7}$$

where f is a real valued function on 3-dimensional space $\mathbb{R}^3$ and C is a constant w.r.t. x,y and z. Surface (7) will be called semi-universal for IR if this constant depends on G (as in (5)) or on (its dual interpretation, see e.g. Egghe and Rousseau (1997))

$$G' = \frac{|ret|}{|\Omega|} , \qquad (8)$$

the fraction of retrieved documents (describing the degree of generality of the retrieval result). This means that, as in (6), any triplet (x,y,z), given G (or G′) fixed, belongs to this surface and, hence, any two such triplets can be compared (on this surface). If this constant is independent of G and G′, hence if C is a universal constant (independent on the IR problem or IR system) we say that (7) is a universal surface for IR.

In this paper we want to discuss the following question :

Question : Can one construct a surface such that any two IR results (independent of a problem or of the IR system) remain on this surface? In short, can one construct a universal surface for IR?

We firstly note that the measures P, R and F only use the sets ret, rel, ¬rel and also in G only rel is appearing. Hence ¬ret, the set of not-retrieved documents is not taken into account. Already for purely principal or formal reasons, this is not acceptable. The dichotomous situation rel/¬rel versus ret/¬ret gives rise to a 2x2 contingency table, hence 4 cells from which only 3 are used in the above measures. Using all 4 cells is similar to the approach in Heaps (1978) where, for the determination of the entropy of a retrieval system, one also uses the information (entropy) coming from all 4 cells as mentioned above.

Another (principal) reason why also ¬ret should be used (if one also uses ¬rel) is the duality principle in IR as explained in Egghe and Rousseau (1997), namely the duality between documents and queries, between indexing and search formulation, between rel and ret. In the vector space model (see e.g. Salton and Mc Gill (1987)) there is even equality between documents and queries (namely a vector with a certain fixed number of coordinates).

These remarks and the fact that the effectiveness surface (6) has the constant G/(1-G), which is only referring to the problem (to rel and not to ret) shows that this theory should

be completed. This will be done in the next section, where the "missing" measure M (miss) will be (re-)introduced and where its properties are proved. "Re-introduced" refers to the fact that in LISA as well as in ISA no English language publication defines this measure and that the "origin" of the use of this measure in a Danish, French and two Russian articles (all of them more than 25 years ago) is not traceable. We will find 3 new efficiency surfaces, hence 4 together with (6). All of these surfaces are not completely universal : two of them (as (6)) depend on the problem (on rel) and the other two depend on the IR action (on ret) so that only IR results can be compared if $|rel|$ is constant (first two surfaces) or if $|ret|$ is fixed (the other two surfaces).

We however deduct from these ("semi-universal") surfaces a totally universal surface, where the four measures P, R, F and M are involved (and nothing else). In other words, any two IR results (no matter what rel or ret is) are on this surface and hence can be compared. We will show that this surface has the equation

$$\frac{P}{1-P} \cdot \frac{1-R}{R} \cdot \frac{F}{1-F} \cdot \frac{1-M}{M} = 1 \qquad (9)$$

and we call it the universal IR surface. This finishes the next section.

The third section studies properties of this surface exploiting the universality of it by defining a uniform distance between any search result (P, R, F, M) and the vector (in $\mathbb{R}^4$) representing the perfect search : (1, 1, 0, 0). Properties of this distance are given and the paper closes with some examples and an open problem.

II. The "missing" measure miss (M) and the universal IR surface

II.1 The measure M

In order to complete the measures derivable from the 2x2 contingence table (ret/$\neg$ret versus rel/$\neg$rel) we write down all four possibilities : $|rel \cap ret|$, $|\neg rel \cap ret|$, $|rel \cap \neg ret|$, $|\neg rel \cap \neg ret|$. We can generally denote this by $|A \cap B|$. In each case we derive

relative measures, by dividing by $|A|$ or $|B|$. This leads to the following complete set of 8 measures (being nothing else than P, 1-P, R, 1-R, F, 1-F and the new measure M and its complement 1-M): see (1), (2), (3) and then

$$M = \frac{|rel \bigcap \neg ret|}{|\neg ret|} \tag{10}$$

$$1-P = \frac{|\neg rel \bigcap ret|}{|ret|} \tag{11}$$

$$1-R = \frac{|rel \bigcap \neg ret|}{|rel|} \tag{12}$$

$$1-F = \frac{|\neg rel \bigcap \neg ret|}{|\neg rel|} \tag{13}$$

$$1-M = \frac{|\neg rel \bigcap \neg ret|}{|\neg ret|} \tag{14}$$

The measure M is the fraction of not-retrieved documents that are relevant, hence it measures the relative quantity of the missed documents. Therefore we call this measure miss.

We looked into the monograph literature, searching for measures, involving M (or at least) $\neg$ret. We failed. We present the following overview. As mentioned, in Dominich (2001), only P, R, F (and G) is used. Boyce, Meadow and Kraft (1995) (from p. 180 on) are in search for other measures than P, R and F but only give 1-P, 1-R and 1-F (respectively called N (noise), O (omission factor) and S (specificity)). Losee (1998), Tague-Sutcliffe (1995), Van Rijsbergen (1979) and Salton and Mc Gill (1987) only use P, R, F (and G) and Frants, Shapiro and Voiskunskii (1997), Losee (1990) and Heaps (1978) only use P and R.

In Grossman and Frieder (1998) other very interesting subjects than evaluation measures are discussed.

Many of the above mentioned books also use (or define) derived measures (mainly from P and R) such as the measure of Dice, but this does not help in the coverage of ¬ret, of course.

As to the journal literature, we performed a LISA and ISA search. LISA contains one Danish article (von Cotta Schoenberg (1976)) where the measure M is called "silence" (there fallout F is called "noise") and ISA contains the articles Levery (1968), Shneiderman (1969), Logunov and Shneiderman (1969) and Pushkarskaya (1968) also using "silence". We were not able to trace the origin of this measure and we are, in any case, to the best of our knowledge, surprised that no reference to "silence" or "miss" is given the last 25 years and that we even lack one reference in English (American) language (except for Shneiderman (1969) being a translation of Logunov and Shneiderman (1969)) ! For this reason and also, as mentioned by one of the referees, since sometimes "silence" is used for the measure 1-R, we will not use this ambiguous name and keep on using the term "miss".

Of course, M cannot be derived directly from an IR result but this is the same problem with F and R (only P is directly calculable). It is well-known (see e.g. Egghe and Rousseau (2001)) that R (in fact rel) can be determined using statistical sampling (yielding also the necessary confidence intervals for the fraction of relevant documents) in which case also F and M are known. So the new measure M is not more complicated than F or R. It is clear from the above that M is the "missing" measure. We also feel that it is a measure at least as interesting as F since, in M, we are talking about relevant documents that are missed (F only deals with non-relevant, retrieved documents).

Since the effectiveness surface of IR (6) contains the 3 measures P, R and F it is intuitively clear that, with the 4 measures P, R, F and M we should be able to determine $\binom{4}{3} = 4$ IR-surfaces (including (6)). This will be done in the next subsection.

II.2 Four semi-universal IR surfaces

We have the following easy result.

Theorem II.2.1 : Let

$$G = \frac{|rel|}{|\Omega|} \tag{15}$$

$$G' = \frac{|ret|}{|\Omega|} \tag{16}$$

Then we have the following 4 surfaces

$$\frac{PF}{R(1-P)} = \frac{(1-F)M}{(1-R)(1-M)} = \frac{G}{1-G} \tag{17}$$

$$\frac{MR}{P(1-R)} = \frac{(1-M)F}{(1-P)(1-F)} = \frac{G'}{1-G'} . \tag{18}$$

The proofs of the four equations are essentially the same. The first equation is already known and the proof is elementary (using elementary set theory).

Note II.2.2.

These 4 surfaces are clearly defined for all values of P, R, F, M $\in$]0,1[. Hence, by continuous extension, these surfaces can be considered for the values P, R, F, M $\in$ [0,1]. Note that for the values P, R, F, M $\in \{0,1\}$ we can easily define this extension since, in all cases we arrive at the undetermined form $\frac{0}{0}$. This follows generally from the above theorem. Let us examine the first equation in (17). Indeed, if P=0 then also R=0 (and vice-versa) and

$$\frac{PF}{R(1-P)} \tag{19}$$

takes the form $\frac{0}{0}$ and hence can be defined (as a continuous extension of the surface (17)) to be $\frac{c}{1-c}$ for all $c \neq 1$ (what we suppose). If $F=0$ then $P=1$ necessarily and again (19) is of the form $\frac{0}{0}$. If finally, $P=1$ then $F=0$ yielding again $\frac{0}{0}$ for (19). This shows that the same is true for the other 3 surfaces in (17) and (18). From now on we will consider these surfaces on the values $P, R, F, M \in [0,1]$.

Note II.2.3.

It is important to remark that all these IR surfaces are "semi-universal" in the sense defined in the introduction. Indeed the two surfaces (17) only depend on G, the generality of the problem (the input, say) and they are independent of the IR process (system, command,...). So on these surfaces, keeping G constant (say we keep the same problem) different IR-results can be compared by using the appearing 3 measures ((P,R,F) or (R,F,M)). The two surfaces (18), on the other hand, only depend on G', the generality of the retrieval result (the output). Again, if G' remains constant, different situations can be compared by using the appearing 3 measures (P,R,M) or (P,F,M). Of course, in (17) one cannot compare the 3 parameter's values if G is not the same and the same for (18) w.r.t. G'. That explains the semi-universality.

It is now clear how to construct a universal IR surface.

Corollary II.2.4.

Independent of the given problem or IR system or command (hence independent of G or G') we have that the measures (P,R,F,M) always belong to the surface (in $\mathbb{R}^4$) :

$$\frac{P}{1-P} \cdot \frac{1-R}{R} \cdot \frac{F}{1-F} \cdot \frac{1-M}{M} = 1 \qquad (20)$$

Proof : This follows directly from (17) or (18). □

The above theorem and corollary show the value of the miss measure M.

Remark II.2.5.

A simpler surface, using all 4 measures P,R,F,M can be given but this is not a universal surface : it contains the values $|ret|$, $|rel|$, $|\neg ret|$, $|\neg rel|$:

$$\frac{P(1-F)}{R(1-M)} = \frac{|rel|\ |\neg ret|}{|ret|\ |\neg rel|} \tag{21}$$

The result follows readily from (1), (2), (13) and (14). Result (21) is interesting but cannot be used in universal IR comparisons because the surface is dependent on both the problem (rel) and the IR result (ret). Surface (20) has the property that all cases of P, R, F, M are on this surface, independent of the problem or the IR result. This means that any two such situations are universally comparable. This is universal IR evaluation which will be explained in the next section.

III. Universal IR evaluation.

Note that also for the universal IR surface (20) we can take $(P,R,F,M) \in [0,1]^4$, 0 and 1 included by continuous extension of the surface on $]0,1[^4$, as was also the case for the surfaces (17) and (18). For this reason, the perfect IR result, being $(P,R,F,M)=(1,1,0,0)$ belongs to the surface (20) since the left hand side gives $\frac{0}{0}$ and hence can be defined to be 1 as a continuous extension.

It is now clear that, since any quadruple (P,R,F,M) belongs to this surface, we can measure the square of the distance between (P,R,F,M) and the vector (1,1,0,0) of the perfect situation.

Since the "worst" result, (0,0,1,1) can occur (at least theoretically) the maximum value of this squared distance is 4. Therefore we define the normalized square of the distance of (P,R,F,M) to (1,1,0,0) as

$$d^2 = \frac{1}{4}\left[(1-P)^2 + (1-R)^2 + F^2 + M^2\right]. \tag{22}$$

d itself is then, of course, the normalized distance. Since $d \in [0,1]$, the measure

$$s = 1 - \frac{1}{2}\sqrt{(1-P)^2 + (1-R)^2 + F^2 + M^2} \qquad (23)$$

is a normalized similarity measure and hence measures how close (P,R,F,M) is to the perfect result (1,1,0,0). Note again that all similarity measures are universal and hence comparable. All aspects of the problem and of the IR result are taken into account in the above measures : ret, rel, $\neg$ret, $\neg$rel, their mutual intersections and even the size $|\Omega|$ of the database.

Note the natural properties that s increases with P and R and decreases with F and M. Of course, for d the opposite is true.

On our way to a theorem on d and s, we formulate the following definitions.

Definitions III.1.

Let $\mathscr{IR}$ represent a certain IR situation, i.e. given by ret, rel, $\neg$ret, $\neg$rel in a fixed database Ω.

1. We say that $\mathscr{D}(\mathscr{IR})$ is the dual IR situation of $\mathscr{IR}$ if their sets $\mathscr{D}$ret of retrieved documents, $\mathscr{D}$rel of relevant documents, $\mathscr{D}\neg$ret of not retrieved documents and $\mathscr{D}\neg$rel of non relevant documents are given by : $\mathscr{D}$ret=rel, $\mathscr{D}$rel=ret.
2. We say that $\mathscr{C}(\mathscr{IR})$ is the complementary IR situation of $\mathscr{IR}$ if their sets $\mathscr{C}$ret of retrieved documents, $\mathscr{C}$rel of relevant documents, $\mathscr{C}\neg$ret of not retrieved documents and $\mathscr{C}\neg$rel of non relevant documents are given by : $\mathscr{C}$ret=$\neg$ret, $\mathscr{C}$rel=$\neg$rel.

Note that $\mathscr{DC}(\mathscr{IR}) = \mathscr{CD}(\mathscr{IR})$ as is easily seen. We have the following theorem.

Theorem III.2.

Let s and d as in (22) and (23) for $\mathscr{IR}$ with P,R,F,M defined via ret, rel, ¬ret, ¬rel. Denote by $s_{\mathscr{D}}$ and $d_{\mathscr{D}}$ the corresponding similarity and distance measure of $\mathscr{D}(\mathscr{IR})$, by $s_{\mathscr{C}}$ and $d_{\mathscr{C}}$ the corresponding similarity and distance measure of $\mathscr{C}(\mathscr{IR})$ and by $s_{\mathscr{DC}}$ and $d_{\mathscr{DC}}$ the corresponding similarity and distance measure of $\mathscr{DC}(\mathscr{IR}) = \mathscr{CD}(\mathscr{IR})$. Then :

$$s = s_{\mathscr{D}} = s_{\mathscr{C}} = s_{\mathscr{DC}} \tag{24}$$

$$d = d_{\mathscr{D}} = d_{\mathscr{C}} = d_{\mathscr{DC}} \tag{25}$$

Proof : If we denote by $\mathscr{D}$(P), $\mathscr{D}$(R), $\mathscr{D}$(F) and $\mathscr{D}$(M) the corresponding Precision, Recall, Fallout and Miss measures of $\mathscr{D}(\mathscr{IR})$ it is easy to see that $\mathscr{D}(P)=R$, $\mathscr{D}(R)=P$, $\mathscr{D}(F)=M$ and $\mathscr{D}(M)=F$. Hence, by (22), $d=d_{\mathscr{D}}$ and hence also $s=s_{\mathscr{D}}$ by (23).

If we denote by $\mathscr{C}$(P), $\mathscr{C}$(R), $\mathscr{C}$(F) and $\mathscr{C}$(M) the corresponding Precision, Recall, Fallout and Miss measures of $\mathscr{C}(\mathscr{IR})$ it is easy to see that $\mathscr{C}(P)=1-M$, $\mathscr{C}(R)=1-F$, $\mathscr{C}(F)=1-R$ and $\mathscr{C}(M)=1-P$. Hence, by (22), $d=d_{\mathscr{C}}$ and hence also $s=s_{\mathscr{C}}$ by (23).

That $d=d_{\mathscr{DC}}$ and $s=s_{\mathscr{DC}}$ follows by application of the above proved results. □

This theorem shows that the similarity and distance measure fully comply with duality and with complements of sets. In a sense, the 4 IR situations $\mathscr{IR}$, $\mathscr{D}(\mathscr{IR})$, $\mathscr{C}(\mathscr{IR})$ and $\mathscr{DC}(\mathscr{IR})$ are equivalent IR situations, at least from a theoretical point of view. It is also logical that IR-evaluation does not evaluate the type of problem given to the system (as expressed by rel, $\mathscr{D}$rel, $\mathscr{C}$ rel, $\mathscr{DC}$ rel) but the combination of the problem with the IR result, in connection with $|\Omega|$. In this sense, the above theorem is natural.

Examples III.3.

1. $|\Omega| = 10^5$ (example: a documentary system in a relatively small discipline such as mathematics), $|rel| = 100$, $|ret| = 200$, $|ret \cap rel| = 50$. Then $P=\frac{1}{4}$, $R=\frac{1}{2}$,

$F=\frac{150}{99,900}$, $M=\frac{50}{99,800}$ yielding d=0.4506946 and s=0.5493054 for the distance, resp. similarity of (P,R,F,M) versus (1,1,0,0).

2. This example is identical with the first one except for Ω : $|\Omega|=10^3$. Now d=0.4594018 and s=0.5405982. Notice the small influence of $|\Omega|$ when the rest remains the same. But example 2 is less similar to (1,1,0,0) due to the fact that the absolute fallout (150) and the absolute miss (50) are the same as in example 1 but in a smaller database.

3. $|\Omega|=10^4$, $|rel|=500$, $|ret|=100$, $|ret \cap rel|=50$. Now we have d=0.5152897 and s=0.4847103. Although $|\Omega|$ is the same as in example 2 and the same is true with $|ret \cap rel|=50$, this IR situation is less similar to (1,1,0,0) (than example 2) due to the high values of $|rel|$ and $|ret|$ (not yielding a higher value of $|ret \cap rel|$).

4. The following examples were suggested by one of the referees (with thanks). It gives 6 rather extreme IR cases describable as follows in 4 categories (see Table 1).

$\|\Omega\|=10^3$	$\|ret\|=10$	$\|ret\|=200$	$\|ret\|=10^3$
$\|rel\|=10$	-	(1,2)	(1,3)
$\|rel\|=200$	(2,1)	-	(2,3)
$\|rel\|=10^3$	(3,1)	(3,2)	-

Table 1. Example of 6 different IR cases of $|rel|$ versus $|ret|$

C_1 : Cases (1,2) and (1,3): the system does not filter enough

C_2 : Cases (2,1) and (3,1): the system filters too much

C_3 : Case (2,3) : non-filtering system

C_4 : Case (3,2) : widest possible problem

Again, notice the duality between C_1 and C_2 and between C_3 and C_4 (cf. Egghe and Rousseau (1997)). We have the following values for d and s in these 6 cases, assuming $rel \cap ret = rel$ in case $|rel| < |ret|$ and $rel \cap ret = ret$ in case $|ret| < |rel|$.

$$d_{(1,2)} = d_{(2,1)} = 0.484596$$
$$s_{(1,2)} = s_{(2,1)} = 0.515404$$
$$d_{(1,3)} = d_{(3,1)} = 0.7035801$$
$$s_{(1,3)} = s_{(3,1)} = 0.2964199$$
$$d_{(2,3)} = d_{(3,2)} = 0.6403124$$
$$s_{(2,3)} = s_{(3,2)} = 0.3596876$$

Note the equality of the measures in case of dual systems, in agreement with Theorem III.2. Notice the weakest performance of the most extreme cases in the table: (1,3) and (3,1) (also situated in the corners of the table, expressing the largest difference between $|\text{rel}|$ and $|\text{ret}|$).

5. We close with a theoretical example : random retrieval. Let $|\Omega|$, $|\text{rel}| = \ell$, $|\text{ret}| = t$ be given. Random retrieval means that we take a random sample of size t in Ω and each document in this sample has a chance of $\frac{\ell}{|\Omega|}$ to be relevant. Consequently, random retrieval is expressed by

$$|\text{ret} \cap \text{rel}| = \frac{\ell t}{|\Omega|} . \tag{26}$$

We have the following results :

$$R = \frac{\frac{\ell t}{|\Omega|}}{\ell} = \frac{t}{|\Omega|} \tag{27}$$

$$P = \frac{\frac{\ell t}{|\Omega|}}{t} = \frac{\ell}{|\Omega|} \tag{28}$$

$$F = \frac{t - \frac{\ell t}{|\Omega|}}{|\Omega| - \ell} = \frac{t}{|\Omega|} = R \tag{29}$$

$$M = \frac{\ell - \frac{\ell t}{|\Omega|}}{|\Omega| - t} = \frac{\ell}{|\Omega|} = P . \tag{30}$$

From this it follows that

$$d^2 = \frac{1}{4}\left[\left(1-\frac{t}{|\Omega|}\right)^2 + \left(1-\frac{\ell}{|\Omega|}\right)^2 + \left(\frac{t}{|\Omega|}\right)^2 + \left(\frac{\ell}{|\Omega|}\right)^2\right] \qquad (31)$$

for such a random retrieval.

Problem III.4

Formula (22) expresses the (square of the) distance of (P,R,F,M) to (1,1,0,0) both on the universal IR surface (20). It expresses, of course, the linear distance between the two points and not the geodetic distance of (P,R,0,0) to (1,1,0,0) over the surface (20), i.e. the length of the shortest curve between (P,R,F,M) and (1,1,0,0) on surface (20). It would be interesting to investigate the properties of such a distance, where one does not "leave" the surface (20), which is the case for any IR result.

IV. Conclusions.

We (re-)introduced the measure miss (M) completing the set of evaluation measures P, R, F and M. We show that no 3 of these measures form a universal IR surface (they form a surface but universality is only obtained if $|ret|$ or $|rel|$ is kept constant, which is almost never true).

We show, however, that the four measures P, R, F, M together form the universal IR surface (20), yielding also the possibility to compare any two IR results, since any obtained set (R,P,F,M) belongs to this surface.

This universal IR evaluation technique is then exploited by considering the distance between any (P,R,F,M) and the vector (1,1,0,0) of the perfect search. If this distance d is normalized then the measure $s=1-d$ yields a similarity measure between (P,R,F,M) and (1,1,0,0), being a universal IR evaluation tool.

References

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

S. Dominich (2001). Mathematical Foundations of Information Retrieval. Kluwer Academic Publishers, Dordrecht, 2001.

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

L. Egghe and R. Rousseau (2001). Elementary Statistics for effective Library and Information Service Management. Aslib-IMI, London (UK), 2001.

V.I. Frants, J. Shapiro and V.G. Voiskunskii (1997). Automated Information Retrieval. Theory and Methods. Academic Press, New York, 1997.

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

H.S. Heaps (1978). Information Retrieval. Computational and theoretical Aspects. Academic Press, New York, 1978.

F. Levery (1968). Role et constitution du thesaurus. Documentaliste, 3, 3-13, 1968.

A.V. Logunov and Y.A. Shneiderman (1969). Experimental evaluation of an sdi system match criterion on a besm-2m computer. Nauchno-tecknicheskaya Informatsiya, Series 2, 2(3), 5-12, 1969 (in Russian).

R.M. Losee (1990). The Science of Information. Academic Press, New York, 1990.

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

R.I. Pushkarskaya (1968). Logic and language of a descriptor system for automatic subject index compilation. Nauchno-tecknicheskaya Informatsiya, Series 2, 2(2), 12-15, 1968 (in Russian).

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

Y.A. Shneiderman (1969). Experimental evaluation of relevance in an sdi system on a besm-2m computer. In : All-Union Institute of Scientific and Research Information. Scientific-Technical Information, Serie 2, n° 3, 5-12, 1969.

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

C.J. Van Rijsbergen (1979). Information Retrieval. Butterworths, London (UK), 1979.

M. von Cotta Schoenberg (1976). Syntactic versus non-syntactic indexing languages. Bogens-Verden, 58(6), 293-294, 1976 (in Danish).