Restricted Evaluation in Information Retrieval

P. Bollmann, Technische Universität Berlin
V. S. Cherniavsky, Technische Universität Braunschweig

1 Introduction

One of the central problems connected with measuring in information retrieval is the problem to select one of numerous measures proposed for evaluating performance of information retrieval systems. In order to select a measure we have to know its properties and we have to be able to compare it with other measures. This in turn brings us to the necessity to describe measures in some precise and operational way. Measures applied to evaluate systems are in fact directly applicable not to systems but rather to their outputs. Hence properties of some measure are always properties exhibited in measuring sets of retrieval outputs and depend on how these sets are selected. The following two approaches are possible here: With the first one we do not limit the retrieval outputs when applying the measure. Within the second approach only those retrieval outputs are permitted which fulfil specific restricting requirements (/1/).

As an example of such a restriction, which is moreover very popular, it may be required that only outputs with the same generality are meaningfully used for measurement. Depending on the approach adopted, the properties of performance measures as well as the results of their comparison may vary drastically. In this paper evaluations of retrieval performance within this second approach are called "restricted evaluations".

There are many measures suggested for evaluating performance of retrieval systems. Some of the most popular ones are recall and precision, recall and fallout, recall-precision-graph, recall-fallout-graph and normalized recall. It is important that these measures represent quite different ideas of being better and there are many problems of interest and importance connected with this fact. It is, for example, both important and interesting that recall and precision for every rank on the one hand and recall-precision-graph on the other are incompatible with respect to the ideas of being better, represented by these measures.

The general formulation of the problem is found in /2/ , special developments and investigations are found for example in /3/ .

However, the results of investigating measures depend radically on wether all possible retrieval outputs are used or not. The above papers (for example /2/ and /3/) are based on unrestricted evaluations. But the approach of unrestricted evaluation encounters severe opposition by some investigators. For example, Robertson assumes in /4/ that in comparing systems the document collections and the total number of relevant documents remain the same.

Some others do use explicitly or implicitly the idea of unrestricted evaluation. Keen (/1/) speaks of "internal" and "external" comparisons in the sense of restricted and unrestricted ones. Rijsbergen (/5/) implicitly allows the comparison of outputs with different generalities by considering all recall-precision pairs. Heine (/6/) intends his measure for comparing all kinds of retrieval outputs.

How essentially restrictions of the above type may influence the properties of measures is shown by the following example. In the investigation /7/ which has already been discussed in /8/ two systems are being compared in order to answer the question of which one is better. To this end a document collection with N documents was fixed as well as a set of t requests. For every request each system yields a distribution of the set of documents over a fixed number of N ranks. This way in /7/ two sequences of distributions

$$\Delta_1, \ldots, \Delta_t$$

$$\Delta'_1, \ldots, \Delta'_t$$

are obtained, where the first line is yielded by the first system and the second one by the second system respectively. Applying a measure μ two sequences of real numbers

$$\mu(\Delta_1), \ldots, \mu(\Delta_t)$$

$$\mu(\Delta'_1), \ldots, \mu(\Delta'_t)$$

are obtained.

This in turn yields consequently a sequence of differences

$$d_i = \mu(\Delta_i) - \mu(\Delta_i') \qquad i = 1, \ldots, t$$

and a sequence of their signs. This last sequence is used in a statistical sign test. Thus measure μ is being concidered only as an ordinal scale (/9/) which means that only its idea of being better is used in the subsequent. In /7/ different measures are used, amongst others the normalized recall R_{norm}:

$$R_{norm} = 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N-n)}$$

the normalized precision P_{norm}:

$$P_{norm} = 1 - \frac{\sum_{i=1}^{n} \ln r_i - \sum_{i=1}^{n} \ln i}{\ln\binom{N}{n}},$$

the Rank-Recall:

$$\text{Rank-Recall} = \frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} r_i}$$

and the Log-Precision:

$$\text{Log-Precision} = \frac{\sum_{i=1}^{n} \ln i}{\sum_{i=1}^{n} \ln r_i}.$$

Here r_i is the rank of the i-th relevant document and n is the number of relevant documents.

The different sign sequences obtained by different measures were concatenated to one general sequence. There are different problems connected to this procedure. The one of interest to us is how valid statistically will be the comparison of the two systems. This validity obviously depends on the length of the sign-squence obtained above, however only under the condition that every sign is obtained independently from every other. This in turn assumes that the measures μ applied in the above procedure are independent. And this is not true: In the situation of the above experiment the retrieval outputs Δ_i and Δ_i' to which the measures have been applied where not arbitrary, they have had always the same number N of documents as well as the same number n_i of relevant ones, which means an essential restriction upon the set of distributions being really compared. It is easily seen that in this situation R_{norm} and Rank-Recall will yield always the same sequence of signs. The same way in this situation, P_{norm} and Log-Precision are equivalent. So the joint sequence of signs has been containing one shorter sequence repeated twice.

This example shows how important it is to precisely take into account restrictions of the type discussed above.

In this paper we approach the problem of how properties and mutual relations of performance measures are affected by replacing the unrestricted measurement approach by a restricted one. We establish a mathematical model describing the comparability of retrieval outputs. We make the assumption that comparability is reflexive, symmetric and transitive. Thus comparability defines an equivalence relation (partition) on the set of all retrieval outputs.

These equivalence relations describe unambiguously the corresponding evaluation restrictions. The following specific restrictions are investigated:

- Only retrieval outputs having the same generality are comparable.
- Only retrieval outputs having both the same number of relevant as well as the same number of non-relevant documents are comparable.
- Only retrieval outputs having both the same number of relevant as well as the same number of non-relevant documents and in addition having the same number of documents in each rank are comparable.

It is shown that restricting the comparability defines a homomorphism on the measures with respect to their hierarchy. Finally the so-called micro-evaluation is described and analyzed from the viewpoint of restricted evaluation.

2 Restrictions

We first repeat in short the definitions of notions which we are going to use in this paper and which are given in full in paper /2/ and /3/ .

We start with the notion of a distribution. We assume a proximity function which assigns a real number or some other formal object to each request-document pair and which moreover takes only a finite number of values. These proximity values are being enumerated and the natural numbers used for this purpose are called ranks. In case all possible proximity values are being enumerated, the ranks are called absolute. In case only those values are enumerated which really have been assigned to some document, the ranks are called relative. A distribution is a mapping of the set of documents into the set of ranks. Accordingly, distributions are absolute or relative. Distributions in which every rank contains at most one document are called (absolute or relative) SMART-distributions. The other distributions are called Taube- distributions.

We assume unweighted relevance judgements and consider all relevant documents (respectively, all nonrelevant documents) equivalent to each other. We shall represent relevant documents by a + sign and nonrelevant documents by a - sign. For example the distribution

$$\left(\begin{matrix}+\ +\\ -\end{matrix}\ \middle|\ \begin{matrix}+\\ --\end{matrix}\ \middle|\ ---\right)$$

is a distribution with three ranks, three relevant and six nonrelevant documents. A user who inspects the first two ranks will retrieve three relevant as well as three nonrelevant documents. For relative SMART-distributions we will omit the vertical lines between the ranks and the brackets. For exam-

ple the distribution.

+ + - + - + - - - -

is a relative SMART-distribution.

We shall use the following notation: Let Δ be a distribution. Let A_v and B_v be respectively the number of relevant and nonrelevant documents retrieved up to the v-th rank. In case Δ has p ranks, Δ is completely described by the vector

$$V(\Delta) = ((A_v, B_v))_{v=1, \ldots, p}.$$

In the subsequent we will not distinguish between Δ and $V(\Delta)$. Let D be the set of all distributions. A viewpoint is a quasiorder (reflexive and transitive) on D. For $\Delta, \Delta' \in D$:

$$\Delta \geq_s \Delta'$$

means, that with respect to viewpoint s, Δ is not worse than Δ'. Let F be a set of formal objects (numbers, vectors, curves etc.) and $\geq_t$ a partial order on F. A measure μ is a mapping

$$\mu : D \to F$$

with

$$\Delta \geq_s \Delta' \Rightarrow \mu(\Delta) \geq_t \mu(\Delta'). \qquad (*)$$

In case (*) holds with "$\Leftrightarrow$" instead of "$\Rightarrow$", s is called the maximal viewpoint of μ.

We say viewpoint s_1 is contained in viewpoint s_2, iff

$$\Delta \geq_{s_1} \Delta' \Rightarrow \Delta \geq_{s_2} \Delta'. \qquad (**)$$

As viewpoints may be considered as subsets of DxD, we write $s_1 \subset s_2$ for (**). Analogously we define the notion of being included for measures: we say measure μ_1 is contained in measure μ_2, iff

$$\mu(\Delta) \geq_{t_1} \mu(\Delta') \Rightarrow \mu(\Delta) \geq_{t_2} \mu(\Delta').$$

In case two measures are mutally included in each other, we say they are equivalent. In case none of them is included in the other one, we call them incompatible. With respect to the inclusion relation thus defined viewpoints and measures form partial orders which may be shown in a single diagram (/3/).

On the base of the notions described above the mechanism of restrictions is developed in the following way. We define a relation $P \subset D \times D$ indicating for any two distributions Δ and Δ' whether or not they are allowed for comparison. We require the following properties for any P.

i $(\Delta,\Delta) \in P$ (reflexivity)
Any distribution may be compared with itself.

ii $(\Delta, \Delta') \in P \Leftrightarrow (\Delta',\Delta) \in P$ (symmetry)
If Δ may be compared with Δ' then Δ' may be compared with Δ.

iii $((\Delta, \Delta') \in P$ and $(\Delta',\Delta'') \in P) \Rightarrow (\Delta,\Delta'') \in P$ (transitivity)

This implies that P is an equivalence relation on D i. e. it is a partition of D. Denoting

$$\{K_i\}_{i \in I}$$

(with I some index set), the equivalence classes, we have

$$\bigcup_{i \in I} K_i = D$$

and

$$K_i \cap K_j = \emptyset \quad i \neq j$$

and $(\Delta,\Delta') \in P \Leftrightarrow$ exists i such that $\{\Delta,\Delta'\} \subset K_i$.

Given a viewpoint s, the relation (s,P) with

$$\Delta \geq_{(s,P)} \Delta' \Leftrightarrow ((\Delta,\Delta') \in P \text{ and } \Delta \geq_s \Delta')$$

is called the restriction of s by P. From the required properties of P it follows that (s,P) is reflexive and transitive, hence it is a quasi-order. Thus (s,P) may be considered a new viewpoint. If we consider s a subset of D x D then $(s,P) = s \cap P$.

We define some partitions which we are going to use in this paper.

a) Let $G(\Delta) = \frac{n}{N}$ be the generality of a distribution where N is the number of documents and n the number of relevant documents in the distribution. We divide distributions into classes, assigning two distributions to the same class iff they have the same generality. The partition thus obtained is denoted by P_G.

Example: Let be

Δ = - + - + + + - - - -
Δ' = (+ _ + | - -)
Then $(\Delta, \Delta') \in P_G$.

b) Two distributions are assigned to the same class iff they have the same number of relevant documents as well as the same number of non-relevant documents. This partition is encountered in cases where two systems are being compared on the base of the same set of documents, of the same set of requests and the same relevance judgements. This situation is found, for example, in the "internal" comparison of Keen (/1/). We denote this partition by P_I.

Example: Let be

$\Delta = \left(\begin{smallmatrix} + & + & | & & | & - & + \\ - & & & & & & - \end{smallmatrix}\right)$

$\Delta' = \left(\begin{smallmatrix} + & | & & + & + \\ & & & - & - \end{smallmatrix}\right)$

Then $(\Delta,\Delta') \in P_I$.

c) Any two distributions having no empty ranks (for example all relative distributions) constitute one class and hence are allowed for comparison. Any distribution having empty ranks constitutes itself a separate class. This situation is found in Cooper's evaluation model (/10/). We denote this partition by P_R.

Example: Let be

$\Delta = \left(\begin{smallmatrix} + & + \\ - & \end{smallmatrix} \mid \; \mid \; - \; - \; \begin{smallmatrix} + \\ - \end{smallmatrix} \; - \right)$

$\Delta' = \quad + - + - -$

Then $(\Delta,\Delta') \varepsilon P_R$ does not hold.

d) Partition P_E is defined by the condition

$(\Delta,\Delta') \varepsilon P_E \iff (p=p' \text{ and } A_v+B_v=A'_v+B'_v \text{ for } v \leq p)$,

where p is the number of ranks of a distribution and A_v is the number of relevant documents and B_v is the number of nonrelevant ones, both up to v-s rank.
Obviously in this case we have for every class K_i a function g_i such that for every v $A_v + B_v = g_i(v)$ holds.
In case K_i contains relative SMART-distributions (exactly one document in every rank) we have $g_i(v) = v, v \leq p$.

Example: Let be

$\Delta = \left(\begin{smallmatrix} + & \\ - & - \end{smallmatrix} \mid \; \begin{smallmatrix} + \\ - \end{smallmatrix} \; - \; - \right)$

$\Delta' = \left(\begin{smallmatrix} + & + \\ - & \end{smallmatrix} \mid \; \begin{smallmatrix} + & + \\ - & - \end{smallmatrix} \right)$

Then $(\Delta,\Delta') \varepsilon P_E$ holds.

e) Using intersections we may obtain new partitions from given ones. In particular we will pay special attention to the partition $P_I \cap P_R \cap P_E$. This partition corresponds to the situation of the evaluation methodology in /7/ described at the beginning of our paper.

Example: Let be

$\Delta = \left(\begin{smallmatrix} + & + \\ - & \end{smallmatrix} \mid \; - \; \mid \; \begin{smallmatrix} + \\ - \end{smallmatrix} \; - \; - \right)$

$\Delta' = \left(\begin{smallmatrix} + \\ - \end{smallmatrix} \; - \mid \; + \; \mid \; \begin{smallmatrix} + \\ - \end{smallmatrix} \; - \; - \right)$

Then $(\Delta,\Delta') \varepsilon P_I \cap P_R \cap P_E$ holds.

Restrictions of viewpoints may be ordered with respect to inclusion. Let s_1 and s_2 be two viewpoints and let P be a partition of D. Then

$$s_1 \subset s_2 \Rightarrow (s_1, P) \subset (s_2, P)$$

holds, that is, restrictions of viewpoints are homomorphisms preserving inclusion.

Analogously measures are ordered on partitions. Let μ_1 and μ_2 be two measures and let P be a partition of D. We say μ_1 is <u>contained</u> in μ_2 on P iff, for any $(\Delta,\Delta') \varepsilon P$,

$$\mu_1(\Delta) \geq \mu_1(\Delta') \Rightarrow \mu_2(\Delta') \geq \mu_2(\Delta')$$

holds. In case μ_1 is contained in μ_2 on P and μ_2 is contained in μ_1 on P, we say μ_1 and μ_2 are <u>equivalent</u> on P.

As usual, we may also order partitions with respect to be finer. The following holds: P_1 is finer than P_2 iff $P_1 \subseteq P_2$. This means that partition P_1 is more restrictive. Then the following holds too:
If

$$(s_1,P_2) \subset (s_2,P_2) \text{ and } P_1 \subset P_2$$

then

$$(s_1,P_1) \subset (s_2,P_1).$$

Analogously we have for measures: if μ_1 is contained in μ_2 on P_2 and P_1 is finer than P_2, then μ_1 is contained in μ_2 on P_1.

This shows that measures and viewpoints are coinciding more and more as possibilities for comparison are more and more restricted. This means that restrticting is a homomorphism of the containment relation of viewpoints and measures. We will give examples for this in the next chapter.

3. Applications

3.1 Hierarchy of measures under restrictions

In /3/ several measures have been compared and ordered polyhierarchically. We will apply the three partitions

$$(P_I \cap P_R \cap P_E) \subset P_I \subset P_G$$

to these measures in order to show how they gradually merge as partitions become finer. We first specify the measures which will be considerd (cf. /3/):

R_v = recall up to the v-th rank,

F_v = fallout up to the v-th rank,

P_v = precision up to the v-th rank

$\Pi(R,F) = ((R_1,F_1), \ldots, (R_p,F_p))$
recall and fallout for every rank,

$\Pi(R,P) = ((R_1,P_1), \ldots, (R_p,F_p))$
recall and precision for every rank,

$\Gamma(R,F)$ = recall-fallout-graph

$\Gamma(R,P)$ = recall-precision-graph

R_{norm} = normalized recall.

Let s_{min} be the minimal viewpoint as defined in /3/. The polyhierarchy of viewpoints and measures of figure 1 where the measures are identified with their maximal viewpoints was obtained in /3/. Standing below means being contained.

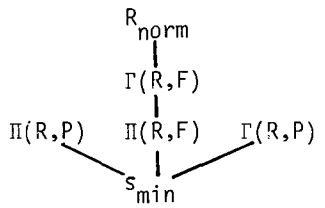

Fig. 1: Polyhierarchy of viewpoints and measures

Let us now consider the effect of restricting.

Restriction by the Partition P_G

From the definition of $\Gamma(R,P)$ (cf. /3/) it follows that $\Gamma(R,F)$ and $\Gamma(R,P)$ are equivalent under this partition. Equation

$$P_v = \frac{G\,R_v}{G\,R_v + (1-G)\,F_v}, \qquad G = \frac{n}{N}$$

implies that $\Pi(R,F)$ is contained in $\Pi(R,P)$ on P_G. As the other relations remain unchanged, we obtain the polyhierarchy of figure 2.

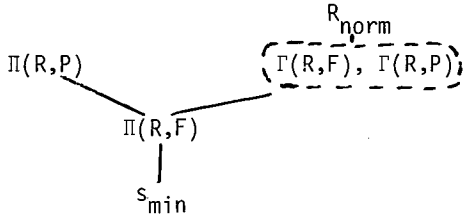

Fig. 2: Polyhierarchy of viewpoints and measures under P_G.

Restriction by the Partition P_I

We show, that in this case $s_{min} \frown P_I$ is the maximal viewpoint for $\Pi(R,F)$.

Let be

$$\Delta = ((A_v,B_v)) \quad v=1, \ldots, p$$

$$\Delta' = ((A'_v,B'_v)) \quad v=1, \ldots, p$$

two distributions, with $n = n'$, $N = N'$ and

$$\Pi(R,F)\ (\Delta) \geq \Pi\ (R,F)\ (\Delta')$$

This implies

$$\left.\begin{aligned} A_v &\geq A'_v \\ B_v &\leq B'_v \end{aligned}\right\} \quad v = 1, \ldots, p$$

Then Δ may be obtained from Δ' by shifting relevant documents to the left and non-relevant documents to the right. This is equivalent to

$$\Delta \geq_{(s_{min},P_I)} \Delta'.$$

All other relations remain unchanged, in particular $\Pi(R,P)$ and $\Gamma(R,P)$ still are incompatible (cf. /3/). We obtain the polyhierarchy given in figure 3.

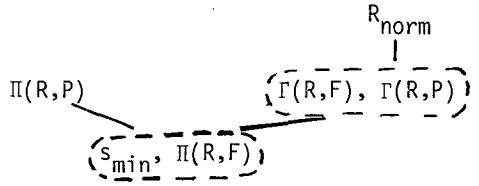

Fig. 3: Polyhierarchy of viewpoints and measures under P_I.

Restriction by the Partition $P_I \frown P_R \frown P_E$

We first show, that in this case R_v, F_v and P_v are equivalent. Let be Δ and Δ' two distributions with $(\Delta,\Delta') \in P_I \frown P_R \frown P_E$.

Then $A_v + B_v = A'_v + B'_v$ holds for $v = 1, \ldots, p$.

$$\begin{aligned} R_v(\Delta) \geq R_v(\Delta') &\iff \\ A_v \geq A'_v &\iff \\ B_v \leq B'_v &\iff \\ F_v(\Delta) \leq F_v(\Delta') & \end{aligned}$$

and

$$\begin{aligned} P_v(\Delta) \geq P_v(\Delta') &\iff \\ \frac{A_v}{A_v + B_v} \geq \frac{A'_v}{A'_v + B'_v} &\iff \\ A_v \geq A'_v &\iff \\ R_v(\Delta) \geq R_v(\Delta'). & \end{aligned}$$

Thus $\Pi(R,P)$ and $\Pi(R,F)$ are shown to be equivalent. Now we are going to show that $\Pi(R,F)$ and $\Gamma(R,F)$ are equivalent too. To this end we show that

$$\Gamma(R,F)\ (\Delta) \geq \Gamma(R,F)\ (\Delta')$$

implies

$$\Pi(R,F)\ (\Delta) \geq \Pi(R,F)\ (\Delta').$$

We prove our assertion by induction on v.

v = 1: Assume $R_1(\Delta) < R_1(\Delta')$. Then $F_1(\Delta) > F_1(\Delta')$ holds. This would imply the situation of figure 4 which contradicts $\Gamma(R,F)\ (\Delta) \geq \Gamma(R,F)\ (\Delta')$. Hence $R_1(\Delta) \geq R_1(\Delta')$ and $F_1(\Delta) \leq F_1(\Delta')$ holds.

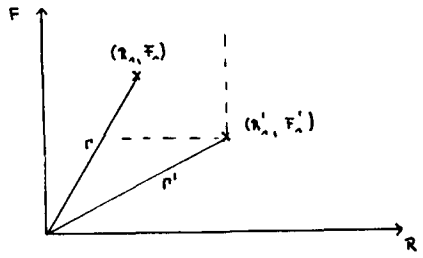

Fig. 4

Assume for v the following holds:

$$R_v(\Delta) \geq R_v\ (\Delta') \quad \text{and} \quad F_v\ (\Delta) \leq F_v\ (\Delta')$$

Then we show that

$R_{v+1}(\Delta) \geq R_{v+1}(\Delta')$ and $F_{v+1}(\Delta) \leq F_{v+1}(\Delta')$

holds too.

Let be $R_{v+1}(\Delta) < R_{v+1}(\Delta')$ and $F_{v+1}(\Delta) > F_{v+1}(\Delta')$ holds.

This would imply the situation of figure 5, which contradicts $\Gamma(R,F)(\Delta) \geq \Gamma(R,F)(\Delta')$.

Hence $R_{v+1}(\Delta) \geq R_{v+1}(\Delta')$ and $F_{v+1}(\Delta) \leq F_{v+1}(\Delta')$ holds.

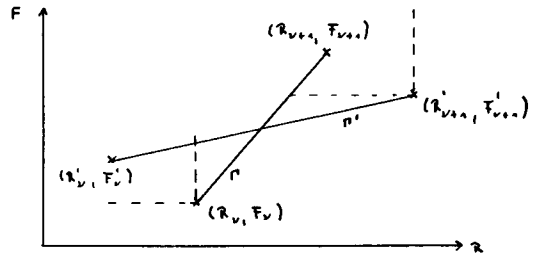

Fig. 5

This way we obtain the hierarchy of viewpoints and measures depicted in figure 6: in this situation all measures are equivalent, except for R_{norm}.

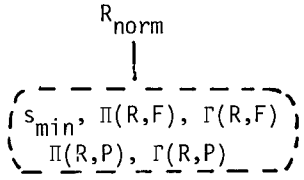

Fig. 6: Hierarchy of viewpoints and measures under $P_I \cap P_R \cap P_E$.

Now moreover consider the measures introduced by Rijsbergen in /5/

$$E_V^{\alpha}(\Delta) = 1 - \frac{1}{\alpha \frac{1}{P_V(\Delta)} + (1-\alpha)\frac{1}{R_V(\Delta)}}$$

$$= 1 - \frac{A_V}{(A_V + B_V) + (1-\alpha)n} \qquad 0 \leq \alpha \leq 1,$$

and a special case of the measure of Pollock (/11/)

$$S_V(\Delta) = \max(R_V(\Delta), P_V(\Delta))$$

$$= A_V \max\left(\frac{1}{n}, \frac{1}{A_V + B_V}\right).$$

We see, that they are all equivalent to R_V under the restriction in consideration, hence R_V, F_V, P_V, E_V^{α} and S_V are equivalent as ordinal scales.

3.2 Micro-evaluation

A widely used and discussed averaging technique is the socalled micro-evaluation. (As far as we know this notion was first used by Rocchio (/12/)). This technique is as follows: Two systems yielding distributions with the same number of ranks are compared on one fixed set of documents and one fixed set of requests, hence we have one distribution for each request and system. (We are not going into details of specific techniques applied in order to make artificially the number of ranks to be equal for all distributions of one system (cf. /13/)). For every system, we merge rankwise all distributions given by this system. This way we obtain one general distribution for each system. These general distributions Δ and Δ' contain the same number of relevant as well as of non-relevant documents, that is $(\Delta,\Delta') \in P_I$. Now a system is evaluated by applying some measure to its general distribution. If in this situation some performance measure has to be selected then this must be done taking into account the polyhierarchy in figure 3. In particular $\Pi(R,P)$ and $\Gamma(R,P)$ are incompatible, $\Gamma(R,F)$ and $\Gamma(R,P)$ are equivalent and $\Pi(R,F)$ is contained in $\Pi(R,P)$. The last relation confirms in this special case Robertson's claim (/14/) that nothing is lost if recall-fallout measures are used instead of recall-precision-measures.

The situation becomes even more restricted in case both systems are assumed to yield SMART-distributions. In this case the above defined distributions Δ and Δ' will have exactly t documents in each rank, if t is the number of requests, that is $(\Delta,\Delta') \in P_I \cap P_R \cap P_E$. In this situation almost all measures are equivalent to each other (see figure 6). Thus it makes no difference what measure is selected as long as only problems of being better are considered.

4 Conclusions

We have shown, that selecting a specific performance measure requires careful identification of the set of objects that are to be measured and compared. Both, a too vast as well as a too narrow identification may lead to distorting results.

References

/1/ E.M. Keen: Evaluation Parameters, in G.Salton (Editor): The SMART Retrieval System, Prentice Hall, Englewood Cliffs, 1971.

/2/ V.S. Cherniavsky; D.G. Lakhuty: Problem of Evaluating Retrieval Systems I, Naucno-Techniceskajy Informazia, Ser. 2, 1970, pp. 24-30 (A.J. 708 141) (Russian), English Translation: Automatic Documentation and Mathematical Linguistics, 4, pp. 9-26.

/3/ P. Bollmann: A Comparison of Evaluation Measures for Document Retrieval Systems, Journal of Informatics, Vol. 1, pp. 97-116, 1977.

/4/ S.E. Robertson: Reply to Bollmann, Journal of Informatics, Vol. 2, pp. 120-121, 1978.

/5/ C.J. van Rijsbergen: Foundations of Evaluation. Journal of Documentation, Vol. 30, pp. 365-373, 1974.

/6/ M.H. Heine: Distance between Sets a an Objective Measure of Retrieval Effectiveness. Information Storage and Retrieval, Vol. 9, pp. 181-198, 1973.

/7/ G. Salton; M. E. Lesk: Computer Evaluation of Indexing and Text Processing, JACM, Vol. 15, pp. 8-36, 1968.

/8/ V.S. Cherniavsky: Problem of Retrieval System Evaluation II: Statistical Evaluation of Retrieval-Systems. About zeros of the sign test. Naucno - Techniceskajy Informazia, Ser. 2, No. 9, 1971 (Russian).

/9/ J. Pfanzagl: Theory of Measurement. Würzburg-Wien, Physika Verlag, 1973.

/10/ W.S. Cooper: Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems. American Documentation, 1968, pp. 30-41.

/11/ S.M. Pollock: Measures for the Comparison of Information Retrieval Systems. American Documentation, 1968, 19, pp. 387-397.

/12/ J.J. Rocchio: Evaluation Viewpoints in Document Retrieval. In: SALTON, G.(ed.) The SMART Retrieval System, Englewood Cliffs, Prentice-Hall, 1971.

/13/ K. Sparck Jones:Automatic Keyword Classification. London: Butterworth, 1971.

/14/ S.S. Robertson: The Parametric Description of Retrieval Tests. Part I: the basic parameters. J. Doc., 1969, 25, pp. 1-27.