Retrieval System Evaluation Using Recall and Precision: Problems and Answers (Extended Abstract)

Vijay V. Raghavan*, Peter Bollmann** and Gwang S. Jung*

*The Center for Advanced Computer Studies,
University of Southwestern Louisiana,P.O. Box 44330, Lafayette, LA 70504
**Technische Universitat Berlin, Fachbereich Informatik, FR 5-11
Franklinstraße 28/29, D-1000 Berlin 10 ,West Germany

1. INTRODUCTION

Retrieval system evaluation plays an important role in judging the efficiency and effectiveness of the retrieval process. Several different evaluation criteria, deemed most critical to user population, were pointed out in[SALTON83, CLEVERDON70]; namely, recall, precision, effort, time, form of presentation and coverage. Among them, recall and precision have received the most attention in the literature. *Recall* is defined as the ratio of the number of relevant documents that are retrieved to the total number of relevant documents. *Precision,* on the other hand, is the number of relevant documents retrieved divided by the number of retrieved documents. In particular, a recall-precision graph is often used as a combined evaluation measure of retrieval systems. Such a graph, given an arbitrary recall point, tells us the corresponding precision value.

Recall and precision are measured after the system determines an ordering on the documents in its collection in response to a user's query. This ordering represents the system's judgement of how well each document relates to the user's need. Based on this, the system can then retrieve items that best suit the user's need, at least, from the system's point of view. Problems arise in *two* situations. The first one occurs when system generates a non-linear ordering of the documents as the output. This implies that system "thinks" two or more items are equally close to the user's search request and would give them identical preference. In this case, some probabilistic notion of precision has to be introduced. A number of measures for this purpose were proposed in the past including, for example, relevance probability and expected precision[COOPER68, COOPER73, YU76, YU77, BOLLMANN88]. We are interested in establishing a correspondence between them and in finding out the extent to which the performance conclusions reached about retrieval systems based on these alternatives agree with each other.

Secondly, when a set of queries is involved and we want to evaluate the overall retrieval results based on this given set of queries, some technique of interpolation of precision values is needed. A method of interpolation based on the use of the *ceiling* operation was utilized in the past[YU77, SALTON71, BUCKLEY85]. We instead propose an interpolation technique which has a nicer interpretation than the precision values obtained by the ceiling method.

In section 2, we give a general introduction to the various concepts and definitions needed in the context of evaluating the retrieval process. In addition, current approaches for measuring recall and precision, as well as problems associated with those are identified. In section 3, alternatives to the existing solutions are advanced and their characteristics are studied. Specifically, in section 3.2, an alternative under the assumption that NR is the stopping criterion is developed. In section 3.3, the implications of using ND as the stopping criterion are considered. In the remainder of section 3, certain important interactions existing between the definition of precision and the choice of stopping criterion are explained. Finally, the conclusions of this study are presented in section 4.

2. BACKGROUND

When a particular search request is presented to a retrieval system, the documents in its collection can be imagined, conceptually, to have been divided into two categories. One consists of the set of relevant documents while the other is the set of nonrelevant ones. In fact, irrespective of what the IR system does, if a document is judged by the user to be of interest, it is *relevant.* It is *nonrelevant,* otherwise. Hence the usefulness of a retrieval system is determined to a great extent by how closely it can imitate the dichotomy identified above.

In order for a retrieval system to locate the relevant items from a given collection with respect to a search request, a measure called the *Retrieval Status Value(RSV)* is often computed between each item in the collection and the search request. The RSV can be viewed as an indicator of the degree of similarity between a document and a request. The RSVs are used to obtain a ranking of items in order that the system can make decisions as to which items should be retrieved.

Two types of ordering of the RSVs can be distinguished immediately: *linear* and *weak* ordering. In the case of a linear or simple ordering, every item in the collection is assigned a distinct RSV by the similarity function used. On the other hand if more than one item may be present at the same level, i.e. with identical RSV, it is termed a weak ordering[BOOKSTEIN76]. In formal terms, a linear ordering is reflexive, transitive, antisymmetric and connected(every pair of elements is comparable). In contrast, a weak ordering may not satisfy antisymmetry[SUPPES57]. In other words, a weak ordering reduces to linear ordering as a special case. Linear ordering greatly simplifies the evaluation of retrieval results in that it imposes a complete constraint on the retrieval order.

Some kind of stopping criterion should be specified for the computation of a pair of recall-precision values. A commonly used criterion is to stop after retrieving a given number of relevant documents. If there are n relevant documents with respect to a given query and assuming that the stopping criterion is the retrieval of h relevant documents, $1 \leqslant h \leqslant n$, there are n possible recall levels, i.e., 1/n, 2/n, . . , h/n, . . , (n−1)/n, and 1.

2.1. Problem of Weak Ordering

Let NR denote the number of relevant documents needed to be retrieved. For a query with n relevant documents, NR ranges between 0 and n. When the ordering produced by the similarity function is linear, for any recall point NR/n, precision is simply calculated as NR/(NR+NNR), where NNR is the number of nonrelevant documents being retrieved along with the NR relevant documents we need. But, when the ordering is not linear, the above method of finding precision must be modified and some notion of probabilistic precision will come into play in the computation. The reason is due to the many possible retrieval orders that may be generated by the system to meet the need. The practice in the past to deal with this situation was the following: Given NR relevant documents to retrieve (corresponding to recall level of NR/n), we start the search from the very top rank, with highest RSV, and keep moving down until we reach a rank where the request can be satisfied. Suppose that there are r relevant documents and i nonrelevant documents at this final rank. It is imagined that the r relevant documents at that rank form r intervals and the i nonrelevant documents at the same rank are uniformly distributed among these r intervals. Hence for every relevant document retrieved, i/r nonrelevant documents is expected to be retrieved[YU76, YU77, SALTON73, SALTON74]. In other words, the total number of nonrelevant documents that are estimated to be retrieved(NNR) is given by

$$NNR = j + \frac{s \cdot i}{r} , \qquad (2.1)$$

where j is the number of nonrelevant documents in ranks completely needed (those above the final rank) and s is number of relevant documents wanted from the final rank. As a result, the precision value at recall level NR/n is defined as

$$\frac{NR}{NR + j + \frac{s \cdot i}{r}} . \qquad (2.2)$$

We refer the evaluation method given in Equation (2.2) as the PRECALL method in the remainder of this paper.

The problem associated with this practice is that the validity of the guess concerning the *typical* distribution of relevant and nonrelevant documents at the final rank is questionable.

This point will be further explained in section 3.2.

2.2. Problem of Multiple Queries

The recall-precision graph is initially defined for a single query. However, in practice, the evaluation result based on a single query is usually not satisfactory. Therefore, many queries are often involved. Since each of the queries might have different number of relevant documents, the *simple* recall levels(i.e., 1/n, 2/n, ..., (n−1)/n and 1) previously introduced can not be used for purposes of averaging, and a method of interpolation of precision values at preselected recall levels is needed.

The conventional choice for these *standardized* recall levels is 0, 0.05, 0.1, ... , 0.95, and 1. The interpolation is done as follows: Each query is processed individually and the precision value with respect to each of the simple recall points is calculated as explained. Following that, the precision values at the various points are scanned in an increasing order, starting from point 2/n. Whenever the precision value being checked at a recall point(say h/n, $h \geqslant 2$) is greater than precision at point (h−1)/n, the precision at point (h−1)/n is changed to the value at point h/n. This can cause a chain-effect. That is, precision at each point k/n ($1 \leqslant k \leqslant h-2$), will also be changed to be the same as the precision value at point h/n, in the event that the precision at k/n is less than precision at point h/n. This whole process is repeated until the last recall point, i.e. 1, has been checked.

After this stage, the precision value corresponding to each of the standardized recall levels (i.e., 0, 0.05, 0.1, ..., 0.95 and 1) is easily determined. Let x be one of the standardized recall levels such that $h/n \leqslant x \leqslant (h+1)/n$ and $0 \leqslant h < n$. Then the precision value at point x is assigned the value at the simple recall point (h+1)/n Since the precision value at the point $x \cdot n$ is the same as that for $\lceil x \cdot n \rceil$, this method is termed the *ceiling* interpolation. As a result, Equation (2.2) will become

$$\frac{\lceil x \cdot n \rceil}{\lceil x \cdot n \rceil + j + \frac{s \cdot i}{r}} . \qquad (2.3)$$

The interpolation process above is performed for each query and the final precision value with respect to each standardized recall is determined by averaging the precision values of all queries at that recall point. Although some other methods of interpolation have been considered in the literature(e.g., [SPARCK-JONES78]), the ceiling method is quite typical of such other methods currently in use.

We refer PRECALL with this ceiling interpolation as the *ceiling*-PRECALL in the remainder of this paper.

2.3. Motivation for Alternative Approaches.

In the remainder of this section we show that the evaluation results obtained using PRECALL are difficult to interpret. We will demonstrate the problem of interpretation by considering the following examples.

Example 2.1 : Suppose we have an ordering

$$\Delta = (+ - - \mid + + + - - - - - - -)$$

There are 13 documents divided into 2 ranks. The first rank consists of 3 documents, one relevant document denoted by + and two nonrelevant documents each of which is denoted by −. The second rank contains 3 relevant and 7 nonrelevant documents. For the recall level .25 the precision value estimated by the PRECALL method is 0.333. □

Some authors claim that precision can instead be represented by P(rel|retr), which is the probability that a retrieved document is relevant. In the next example, we want to illustrate this probability for the recall level 0.25.

Example 2.2 : Let the ordering be same as in example 2.1. The recall level 0.25 corresponds to retrieving one relevant document. Hence the probability that a retrieved document is relevant for the recall level .25 is equal to the probability that a retrieved document is relevant given that we desire one relevant document. There are three possible arrangements of the documents in the first rank each of which have the probability of 1/3 : + − − , − + −, − − +. We have

$$P(\text{rel} \mid \text{retr}) = \frac{P(\text{rel} \cap \text{retr})}{P(\text{retr})} , \qquad (2.4)$$

where

$$P(\text{retr}) = \sum_{v=0}^{2} P(\text{retr} \mid \text{arrangement}_v) P(\text{arrangement}_v).$$

We now obtain

$$P(\text{rel} \cap \text{retr}) = \frac{1}{13} \quad ,$$

since exactly one relevant document is retrieved. For the three arrangements, let arrangement_v mean that v nonrelevant items will be retrieved with that arrangement for getting one relevant item. Then, since v+1 documents are retrieved altogether we get

$$P(\text{retr} \mid \text{arrangement}_v) = \frac{v+1}{13} \quad .$$

It follows that

$$P(\text{retr} \mid \text{arrangement}_0)P(\text{arrangement}_0)$$
$$= \frac{0+1}{13} \cdot \frac{1}{3} = \frac{1}{13} \cdot \frac{1}{3} \quad \text{and}$$
$$P(\text{retr} \mid \text{arrangement}_1)P(\text{arrangement}_1)$$
$$= \frac{1+1}{13} \cdot \frac{1}{3} = \frac{2}{13} \cdot \frac{1}{3} \quad \text{and}$$
$$P(\text{retr} \mid \text{arrangement}_2)P(\text{arrangement}_2)$$
$$= \frac{2+1}{13} \cdot \frac{1}{3} = \frac{3}{13} \cdot \frac{1}{3} \quad .$$

Hence,

$$P(\text{rel} \mid \text{retr}) = 0.5. \qquad \square$$

We refer to $P(\text{rel} \mid \text{retr})$ as the *Probability of Relevance* (PRR) in the remainder of this paper.

Another way to define precision in an average sense is to ask what precision we can expect to obtain for the recall level 0.25. In the next example we consider this alternative, which will be referred to as the *Expected Precision* (EP).

Example 2.3 : Suppose that the ordering is the same as in example 2.1. We ask now what precision we can expect at the recall level 0.25, or equivalently, when we desire one relevant document. Again we have the same three arrangements, as in example 2.2. For the first arrangement precision is 1, for the second it is 0.5 and for the third it is 1/3. Hence, for expected precision, we get

$$EP = 1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{3} = \frac{11}{18} \approx 0.611 \qquad \square$$

We have shown that for a given recall point there are at least three possible definitions of precision. For our example precision could be 0.333 or 0.5 or 0.611 depending on how we define precision in an average sense. Note also that what we call as PRECALL is neither PRR nor EP. Thus, the meaning of PRECALL is hard to explain. Moreover the situation is further complicated by the fact that these precision values can contradict each other. We show that through the following two examples.

Example 2.4 : Let

$$\Delta = (+--|+++-------)$$

as before, and let

$$\Delta' = (+++-----|+----)$$

be another retrieval ordering. We again compute the precision values for the recall level of 0.25 according to the three different definitions:

	PRECALL	PRR	EP
Δ	0.333	0.500	0.611
Δ'	0.375	0.444	0.609

We see that, for recall point 0.25, Δ' is better than Δ when PRECALL is used but Δ is better than Δ' for PRR and EP. $\square$.

In the following example we want to show that PRR and EP can also contradict at a given recall point.

Example 2.5 : Let Δ and Δ' be as follows:

$$\Delta = (+- \mid +++++------ \mid ++++----)$$
$$\Delta' = (++++++---- \mid ++-- \mid ++----)$$

For the recall level 0.1, or equivalently for retrieving one relevant document, we obtain precision values for those two definitions as

	PRR	EP
Δ	0.667	0.750
Δ	0.636	0.775

$\square$

From the examples presented in this section we see that, in the case that the retrieval output is a weak ordering, there are several ways to define precision for a given simple recall point. Depen-

ding on these definitions there are different interpretations associated with the evaluation results given by recall and precision. Furthermore, we believe that there is no simple and intuitively reasonable interpretation of precision values, as a function of recall, obtained by the PRECALL method. In contrast we find that PRR and EP represent reasonable methods for handling weak ordering and are therefore promising alternatives to the PRECALL method. However, since PRECALL has certain historical significance, we should look for ways to interpret those values even if the meaning might be somewhat more convoluted.

3. PROPOSED SOLUTIONS AND THEIR CHARACTERISTICS

3.1. General Concepts

In the previous section we introduced two different methods of computing precision in an average sense mentioned above; namely, Probability of Relevance and Expected Precision. In the developments that follow, each method will be investigated with respect to two distinct stopping criteria, namely the number of relevant documents that are to be retrieved(NR) and the desired number of retrieved documents(ND). Therefore, essentially there are four different possible combinations, i.e., PRR vs. NR or PRR vs. ND or EP vs. NR or EP vs. ND. Other stopping criteria are possible; for example, number of nonrelevant documents that are retrieved(NNR)[KRAFT79]. By the way, it should be noted that there is an immediate correspondence between NR and one of the standardized recall levels described at the end of previous section. Let us suppose there are n relevant documents with respect to a query. Given a standardized recall level x, the corresponding NR is simply x·n. Hence, depending on x and n, NR is not restricted to integer numbers only. For example, let there be 30 relevant documents in response to a query, the 10 predefined NR points are 0, 1.5, 3, ..., 28.5 and 30.

Before discussing the various properties associated with the above measures of evaluation, symbols and notations that are most frequently needed in the remainder of this paper are introduced next. Some others will be explained later as the need arises. Given a search request in terms of the number of relevant documents wanted(NR), the retrieval system begins search from the highest level(rank 1), which by definition contains documents with the highest RSV. It continues until the final level(say rank l_f) at which the stopping criterion is met. We now define the following notations:

t: number of documents searched through in ranks 1 through (l_f-1).

t_r: number of relevant documents searched through in ranks 1 through (l_f-1).

j: number of non-relevant documents searched through in ranks 1 through (l_f-1).

r: number of relevant documents in rank l_f.

i: number of non-relevant documents in rank l_f.

3.2. PRR vs. NR

In section 2 we defined PRR = P(rel | retr) given that the user requires NR relevant documents. For the developments given in this section we let P_v denote the probability that v nonrelevant documents are retrieved in l_f. That is,

$$P_v = \text{Prob}(\text{v nonrel. docs. retrieved in } l_f \mid \text{s rel. docs. retrieved in } l_f) \qquad (3.1)$$

Furthermore, let s denote $NR-t_r$, the number of relevant documents to be retrieved at l_f. Then based on Cooper's definition that

$$esl_{NR} = \sum_{v=0}^{i} (j + v)P_v \ ,$$

the following theorem is obtained.

Theorem 3.1 :

$$P(\text{rel} \mid \text{retr}) = \frac{NR}{NR + esl_{NR}} \qquad (3.2)$$

Proof:

$$PRR = P(\text{rel} \mid \text{retr}) = \frac{P(\text{rel} \mid \text{retr})}{P(\text{retr})} \ .$$

Let N be the number of documents in the collection. Since NR relevant documents are retrieved, we obtain

$$P(\text{rel} \cap \text{retr}) = \frac{NR}{N} \ .$$

If v nonrelevant documents are retrieved in rank l_f then

$$\text{Prob}(\text{retr} \mid \text{v nonrelevant docs. retrieved in } l_f) = \frac{NR + j + v}{N} .$$

Let P_v be as defined in Equation (3.1). Then

$$P(\text{retr}) = \sum_{v=0}^{i} P(\text{retr} \mid \text{v nonrelevant docs. retr. in } l_f) P_v$$

$$= \sum_{v=0}^{i} \frac{NR + j + v}{N} P_v = \frac{NR}{N} + \frac{1}{N} \sum_{v=0}^{i} (j + v) P_v$$

$$= \frac{NR}{N} + \frac{1}{N} esl_{NR} .$$

Hence, we obtain

$$PRR = \frac{\frac{NR}{N}}{\frac{NR}{N} + \frac{1}{N} esl_{NR}} = \frac{NR}{NR + esl_{NR}} \quad \square$$

From Cooper[COOPER68] we know that

$$esl_{NR} = j + \frac{i \cdot s}{r + 1} .$$

From this we finally obtain the closed form expression for PRR.

$$PRR = \frac{NR}{NR + j + \frac{i \cdot s}{r + 1}} \qquad (3.3)$$

Although Cooper derives this expression and interprets it as expected precision, we show here that it is more correctly interpreted as P(rel | retr). Based on Equation (3.3), Cooper points out that, in computing esl vs. NR, the r relevant document should be imagined as forming r+1 intervals. Note however that if we replace r+1 by r, this equation reduces to Equation (2.2). Thus the assumption made, in computing PRECALL, about the distribution of documents in l_f is not consistent with that for PRR. This important observation further strengthens our belief that Equation (2.2) may not be used without further justification.

In section 2, we establish the need for the interpolation of precision values at standardized recall levels, when evaluation is to be performed on the basis of many queries. We also explained the scheme used in the past to cope with such a situation and the problem associated with that scheme. In the remainder of this section, we propose a method of interpolation which is found to be more natural than the ceiling interpolation and the interpolated values still have meaning as a conditional probability.

The idea behind the *intuitive* interpolation originated from the possibility that we can make use of the functional relationship between a set of recall levels and integer values of NR. That is, given a recall level x, the corresponding NR is x·n, where n is the total number of relevant documents in response to the given query. We can therefore determine NR associated with an arbitrary x. Similarly, we can also consider the functional relationship between esl and s values and then make an appropriate substitution in Equation (3.3). Hence, we propose the following expression, for $0 < s \leqslant r$,

$$PRR = \frac{x \cdot n}{x \cdot n + j + \frac{i \cdot s}{r+1}} . \qquad (3.4)$$

Notice that s can be a fractional number with this modification. The above expression is next formally justified by generalizing Cooper's closed form formula for esl, for real values of s. From probability theory we know that

$$P_v = \frac{\frac{(s-1+v)!}{(s-1)!v!} \frac{(r-s+i-v)!}{(r-s)!(i-v)!}}{C_i^{r+i}} .$$

for integer s. If we now interpolate all factorials that contain a s with the Γ function then,

$$esl = j + \frac{i \cdot s}{r + 1} \quad \text{for } 0 < s \leqslant r .$$

With this result we find a simple formula for esl for all values of s and it can be used for computing PRR. It is important to note that, irrespective of whether s is integer, we can show $\sum_{v=0}^{i} P_v = 1$ for all s. Since the P_vs remain as probabilities, PRR continues to have interpretation as a conditional probability.

In the intuitive interpolation we provide a method to deal with possible fractional number s. By the same token we also need to consider a method of extrapolation when s is very small. There are two cases to be considered.

The first situation is when we have at least one relevant document in the first rank. From Equation (3.4)

$$PRR = \frac{s}{s + \frac{s \cdot i}{r + 1}} = \frac{r + 1}{r + i + 1} .$$

Hence s is not involved in the computation of

PRR.

The second case is when we have some ranks that have only nonrelevant documents before the first rank that contains relevant documents. Let $j > 0$ be the number of those nonrelevant documents. Again from Equation (3.4)

$$PRR = \frac{s}{s + j + \frac{s \cdot i}{r + 1}} .$$

Hence,

$$\lim_{s \to 0} PRR = 0 .$$

3.3. PRR vs. ND and EP vs. ND

Following the ideas of section 3.2 PRR vs. ND is defined as

$$PRR = P(rel \mid retr) ,$$

given that the user stops searching after having retrieved ND documents. In relation to this stopping criterion, let the number of documents to be retrieved in l_f in order to meet the stopping criterion be denoted by k. That is, $k = ND - t$. Let μ be the expected number of relevant documents retrieved in l_f. We assume r and i to be as defined in section 3.1. Then

$$\mu = \frac{kr}{r + i} , \qquad (3.5)$$

since μ is the expected value of a hypergeometrically distributed random variable.

Theorem 3.2 :

$$PRR = \frac{1}{ND}\left(t_r + \frac{kr}{r + i}\right) .$$

Proof: Let Q_v denote the conditional probability that v relevant documents are retrieved in l_f given that k documents are retrieved from that rank.

$$PRR = \frac{P(rel \cap retr)}{P(retr)} .$$

$$P(retr) = \frac{ND}{N} .$$

$$P(rel \cap retr)$$

$$= \sum_{v=0}^{r} P(rel \cap retr \mid v \text{ rel. docs. retrieved in } l_f) Q_v$$

$$= \sum_{v=0}^{r} \frac{t_r + v}{N} Q_v = \frac{t_r}{N} + \frac{1}{N} \sum_{v=0}^{r} v Q_v = \frac{t_r}{N} + \frac{1}{N} \cdot \mu .$$

$$P(rel \cap retr) = \frac{t_r}{N} + \frac{1}{N} \cdot \frac{k \cdot r}{i + r} \text{ by Equation 3.5 } \square$$

If the stopping criterion is ND, then EP is defined as

$$EP = \sum_{v=0}^{r} \frac{t_r + v}{ND} Q_v = \frac{\frac{1}{N} \sum_{v=0}^{r} (t_r + v) Q_v}{\frac{1}{N} ND}$$

$$= \frac{P(rel \mid retr)}{P(retr)} = PRR .$$

Hence if ND is the stopping criterion then EP is equal to PRR.

3.4. A parametric Description of the PRECALL-Graph with Intuitive Interpolation.

In section 2.3 we showed that problems of interpretation arise if we consider precision given by PRECALL as a function of recall. Specifically, given a graph as in Fig. 3.1, p may *not* be interpreted as either P(rel | retr) or expected precision corresponding to a recall level of r. Thus, it is still an open problem to explain the meaning of PRECALL. For the examples considered in section 2.3, these problems remain regardless of the type of interpolation used.

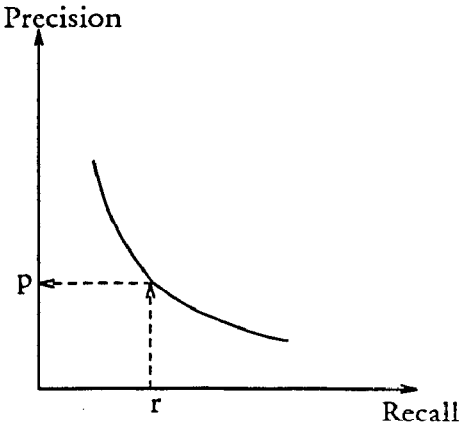

Figure 3.1 : Interpretation of PRECALL as a precision for a given recall.

However we can develop an approach that yields an interpretation of the PRECALL-Graph with intuitive interpolation, by using ND as a common parameter. For the convenience of discussions that follow, this method will be re-

ferred to as *intuitive*-PRECALL. Note that in this method Equation (2.2) is applied regardless of whether or not NR is an integer. In order to develop an interpretation for *intuitive*-PRECALL graph we define P(retr | rel) vs. ND and expected recall (ER) vs. ND analogous to the definitions of PRR and EP.

$$ER = P(retr \mid rel) = \frac{1}{n}(t_r + \frac{kr}{r+i}) \ ,$$

where n is the number of relevant documents in the collection. From Equation (2.2), we know that *intuitive*-PRECALL method yields the points given by the coordinates

$$\left(R \ , \ \frac{nR}{nR + j + \frac{(nR - t_r)i}{r}} \right)$$

for $0 < R \leqslant 1$. We use this relationship and the expression derived for ER in order to establish the connection between EP and ER, under the condition that they are both given as a function of ND.

We now obtain the following interpretation of the PRECALL-Graph: Given any integer ND, for $0 < ND \leqslant N$ and $r \geqslant 1$, there exists a point on the graph obtained for *intuitive*-PRECALL whose coordinates are exactly ER and EP. In other words, the PRECALL-Graph with intuitive interpolation includes every (ER, EP) pair obtainable via ND. Hence one correct way to interpret this graph is given in figure 3.2.

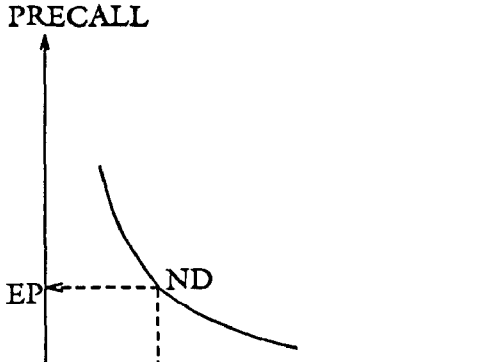

Figure 3.2 : Parametric interpretation of the graph obtained by *intuitive*-PRECALL method.

This interpretation of the *intuitive*-PRECALL graph requires an indirect approach similar to that mentioned in van Rijsbergen[RIJSBERGEN79], where he describes Recall and Precision as a function of a common parameter λ.

The above analysis provides an interpretation of points on the graph obtained by the *intuitive*-PRECALL method for one query. When many queries are involved, the interpretation can easily be extended if the averaging is done over ND. But averaging over NR is still a problem vis-a-vis the meaning that can be given to points on the resulting graph. We have however not been able to find even such an indirect interpretation for the graph obtained by the PRECALL method with ceiling interpolation.

3.5. Precision as a function of Recall, Fallout and Generality

Robertson[ROBERTSON69] showed that

$$\text{Precision} = \frac{\text{Genarality} \times \text{Recall}}{\text{Generality} \times \text{Recall} + (1 - \text{Generality}) \times \text{Fallout}} \ , \qquad (3.6)$$

where Generality G is defined as $G = n/N$. In what follows we want to discuss how the definition of precision as either PRECALL or PRR is compatible with Equation (3.6). First let us consider PRECALL. Let R denote recall and F be fallout. Then the usual Recall-Fallout-Graph is defined by plotting, for every full rank, a Recall-Fallout point into the Recall-Fallout plane and then interpolating these points linearly[ROBERTSON69]. Hence for any recall R we obtain

$$F = \frac{j}{N-n} + \frac{(nR - t_r)i}{(N-n)r} \ . \qquad (3.7)$$

If we substitute Equation (3.7) to (3.6) we obtain

$$\frac{GR}{GR + (1-G)F} = \frac{nR}{nR + j + (nR - t_r)\frac{i}{r}} \ .$$

Since $NR = nR$ and $s = nR - t_r$ we find out that $\frac{GR}{GR + (1-G)F}$ is precisely PRECALL with intuitive interpolation. Hence we can imagine the PRECALL-Graph with intuitive interpolation as a mapping from the traditionally defined Recall-Fallout-Graph given by Equation (3.7). More specifically, given any (R, F) pair, the transformation

$$(R,F) \rightarrow \left(R, \frac{GR}{GR + (1-G)F} \right)$$

yields a point on *intuitive*-PRECALL-Graph and vice-versa. Here PRECALL may have some meaning indirectly through (interpolated) the F values given by the Recall-Fallout-Graph. This depends on whether proper meaning can be given to the interpolated values of fallout, as specified in Equation (3.7). This definition of the Recall-Precision-Graph was proposed by Bollmann[BOLLMANN78].

Thus we see that, for a given recall, Equation (3.6) establishes a different notion of precision depending on how fallout is defined. Furthermore we will face similar problems as in the case of precision if we want to investigate the meaning of fallout, defined in different ways, as a function of recall. In other words, interpolated fallout values given by Equation (3.7) are not interpretable as the probability of retrieving a nonrelevant document or as the expected value of the ratio of NNR to the number of documents retrieved.

4. CONCLUSIONS

Two interesting problems that arise, when using recall and precision as measures of retrieval system performance, are due to the weak ordering of output and the need for handling multiple queries. The seriousness of these problems is also determined by the choice of the stopping criterion (e.g. number of relevant documents retrieved (NR) or number of documents retrieved (ND)).

With respect to the problem of weak ordering, two different notions of probabilistic precision are considered: Probability of Relevance(PRR) and Expected Precision(EP). Although these notions entail the possibility of combinatorial explosion in assessing the various orderings of outputs, it is shown that PRR vs. ND, PRR vs. NR can be handled by relatively efficient computational procedures.

The problem associated with averaging of precision over a number of queries arises only when NR is chosen as the stopping criterion. To handle this problem, a method of interpolation that allows the computation of precision for non-integral values of NR is needed. For PRR vs. NR an interpolation technique is advanced which is natural and has a sound formal justification.

We believe that PRR vs. Recall (or, equivalently, NR) has the advantage of having a well defined meaning. Furthermore, it is closely related to expected search length [COOPER68] and lends itself to efficient computation.

In contrast the *ceiling*-PRECALL, which has been used in many previous studies, is not amenable to any reasonable interpretation. The problem is caused not only by the fact that averaging for multiple queries is done over NR but also by the fact that the method of interpolation is ad hoc. However, we are able to show that *intuitive*-PRECALL method yields a graph that can be given a sound interpretation, if ND is viewed as the parameter through which recall and precision are defined. Thus, our results here suggest that intuitive-PRECALL method, for averaging purposes, should take precision values over many queries at fixed ND(and not NR). Even though *intuitive*-PRECALL method with ND as the stopping criterion gives a sound interpretation, it may have practical difficulties in the selection of NDs as follows. When the number of documents in a collection is very large we must select several NDs for which ERs and EPs are obtained. When ND is incremented by fixed intervals, one may not get desired ER points that covers whole range of possible values(i.e., ER values may be so close together that one may have difficulty to use these as criterion for comparison) since relevant documents are likely to be unevenly distributed among the various ranks. However with well selected NDs one can use this measure very meaningfully. In this paper we also identify the origin of the *intuitive*-PRECALL method and its connection to the Recall-Fallout-Graph defined by Robertson[ROBERTSON69]. The question of how to interpret *intuitive*-PRECALL, with averaging over NR, is yet to be addressed.

With respect to the other measure, EP, we show that EP vs. ND coincides with PRR vs. ND. However the problem of computing EP vs. NR needs to be given a treatment similar to that of PRR vs. NR. More specifically, the equations of how to obtain a closed form formula for EP as well as what is a natural method of interpolation for EP are still being addressed. We will provide some answers in these directions in [BOLLMANN89].

It is hoped that this investigation contributes to a better understanding of precision defined as a function of NR or ND as methods of evaluation and in the systematic selection of techniques to deal with problems of weak ordering and multiple queries.

5. REFERENCES

[BOLLMANN78] Bollmann, P., "A comparison of Evaluation Measures for Document Retrieval System," *Journal of Informatics*, vol. 2, no. 1, 1978.

[BOLLMANN88] Bollmann, P. and Raghavan, V.V., "A utility-theoretic analysis of expected search length," *Proceedings of the 11th International Conference on Research and Development in Information Retrieval*, 1988, Grenoble, France.

[BOLLMANN89] Bollmann, P. Raghavan, V.V., Jung, G.S. and Shu, L. "Probability of Relevance and Expected Precision in Evaluating Retrieval Performance," in preparation.

[BOOKSTEIN76] Bookstein, A. and Cooper, W.S., "A general mathematical model for information retrieval systems," *Library Quarterly*, vol. 46, 1976, pp. 153-157.

[BUCKLEY85] Buckley, C., "Implementation of the SMART Information Retrieval System", Technical Report 85-686, Department of Computer Science, Cornell University. May 1985.

[CLEVERDON70] Cleverdon, C.W., "Evaluation of tests of information retrieval systems," *Journal of Documentation*, vol. 26, 1970, pp. 55-67.

[COOPER68] Cooper, W.S., "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems," *American Documentation*, vol. 19, 1968, pp. 30-41.

[COOPER73] Cooper, W.S., "On selecting a measure of retrieval effectiveness," *Journal of the American Society for Information Science*, vol. 24, 1973, pp. 87-100.

[KRAFT79] Kraft, D.H. and Lee, T., "Stopping Rules and Their Effect on Expected Search Length," *Information Processing & Management*, vol. 15, 1979, pp. 47-58.

[RIJSBERGEN79] van Rijsbergen, C.J., *Information Retrieval*, 2nd. Ed., Butterworths Co., Ltd., 1979.

[ROBERTSON69] Robertson, S.E. "The Parametric Description of Retrieval Tests. Part II: Overall Measures," *Journal of Documentation*, vol. 25, 1969, pp. 93-107.

[SALTON71] Salton, G.(Ed.), *The Smart Retrieval System-Experiments in Automatic Document Processing*, Prentice-Hall, Englewood-cliffs, N.J., 1971.

[SALTON73] Salton, G. and Yang, S.G., "On the Specification of Term Values in Automatic Indexing," *Journal of Documentation*, vol. 29, no. 4, 1973, pp. 351-372.

[SALTON74] Salton, G., Yang, C.S. and Yu, C.T., "Contribution to the Theory of Indexing", *Information Processing 74*, North-Holland, Pub., Co., 1974, pp. 584-590.

[SALTON83] Salton, G., and McGill, M.J., *Introduction to Modern Information Retrieval,* McGraw-Hill, Inc. N.Y., 1983.

[SPARCK-JONES78] Sparck Jones, K., "Performance averaging for recall and precision," *Journal of Informatics*, vol. 2, 1978, pp. 95-105.

[SUPPES57] Suppes, P., *Introduction to Logic*, Van Nostrand, N.Y., 1957.

[YU76] Yu, C.T. and Salton, G. "Precision-weighting-an effective automatic indexing method," *JACM*, vol. 23, 1976, pp. 76-88.

[YU77] Yu, C.T. and Raghavan, V.V., "A single-pass method for determining the semantic relationship between terms," *Journal of the American Society for Information Science*, vol. 28, 1977, pp. 345-354.