# A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance

VIJAY V. RAGHAVAN and GWANG S. JUNG
University of Southwestern Louisiana
and
PETER BOLLMANN
Technische Universitat Berlin

Recall and precision are often used to evaluate the effectiveness of information retrieval systems. They are easy to define if there is a single query and if the retrieval result generated for the query is a linear ordering. However, when the retrieval results are weakly ordered, in the sense that several documents have an identical retrieval status value with respect to a query, some probabilistic notion of precision has to be introduced. Relevance probability, expected precision, and so forth, are some alternatives mentioned in the literature for this purpose. Furthermore, when many queries are to be evaluated and the retrieval results averaged over these queries, some method of interpolation of precision values at certain preselected recall levels is needed. The currently popular approaches for handling both a weak ordering and interpolation are found to be inconsistent, and the results obtained are not easy to interpret. Moreover, in cases where some alternatives are available, no comparative analysis that would facilitate the selection of a particular strategy has been provided. In this paper, we systematically investigate the various problems and issues associated with the use of recall and precision as measures of retrieval system performance. Our motivation is to provide a comparative analysis of methods available for defining precision in a probabilistic sense and to promote a better understanding of the various issues involved in retrieval performance evaluation.

Categories and Subject Descriptors: H.3.0 **[Information Storage and Retrieval]**: General; H.3.3 **[Information Storage and Retrieval]**: Information Search and Retrieval—*retrieval models*; H.3.m **[Information Storage and Retrieval]**: Miscellaneous—*systems evaluation, performance measurement*

General Terms: Experimentation, Measurement, Performance, Theory

Additional Key Words and Phrases: Evaluation measures, expected precision, expected search length, fallout, generality, information retrieval, precision, probability of relevance, recall, stopping criterion

## 1. INTRODUCTION

Retrieval system evaluation plays an important role in judging the efficiency and effectiveness of the retrieval process. Many evaluation methods have been proposed and investigated in the past [2, 3, 8, 11, 13, 14, 17, 20, 21, 24, 28, 30, 31].

Authors' addresses: V. V. Raghavan and G. S. Jung, The Center for Advanced Computer Studies, University of Southwestern Louisiana, P.O. Box 44330, Lafayette, LA 70504; P. Bollmann, Technische Universitat Berlin, Fachbereich Informatik, FR5-11, Franklinstraße 28/29, D-1000, Berlin 10, West Germany.

The most comprehensive and objective approach should take many aspects of the retrieval process into consideration. These include, for example, the resources used by the system to perform a retrieval operation, the amount of time and effort spent by a user to obtain needed information, and the ability of the system to retrieve useful items. This approach is hard to realize, if not impossible, because it is difficult to obtain all the relevant measurement parameters [21, 25]. Even if it were possible to have all the information available, how to combine them appropriately to obtain a single measure is another question. Consequently, it is common practice in research investigations to concentrate mainly on measures pertaining to the quality of the retrieval output.

From the user's point of view, most people would generally agree to the prescription that a retrieval system should behave as follows. "Retrieve as many relevant items as possible and as few nonrelevant items as possible in response to a request." Roughly speaking, the former criterion corresponds to the concept of *Recall*, and the latter one pertains to the notion of *Precision*.

Recall and precision are often conflicting goals in the sense that if one wants to see more relevant items (i.e., to increase recall level), usually more nonrelevant ones are also retrieved (i.e., precision decreases). The converse is also true [10, 15]. The question thus arises as to how we can claim that system A is better than system B only on the basis of determining recall and precision. Traditionally, system A is assumed to be better than system B if, at every recall point, A's precision value is higher than B's. If this does not hold, then the precision values for selected recall values are averaged and compared.

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 judgment of how well each document relates to the user's need. On the basis of this judgment, the system can then retrieve items that receive sufficiently high ranks. Problems arise in *two* situations. The first one occurs when a system generates a weak ordering of the documents as the output. This implies that the 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 [4, 11, 13, 33, 34]. We are interested in establishing a correspondence between these measures and in finding out the extent to which the performance conclusions reached about retrieval systems on the basis of these alternatives agree with each other.

The second problem arises when a set of queries is involved. If 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 [7, 21, 34]. With this method, the interpretation of precision is difficult and not amenable to objective treatment when all the documents in the final rank are not retrieved. We instead propose an interpolation technique that allows the interpolated values to be interpreted identically regardless of whether a rank is fully or partially retrieved. In addition, experimental results that enable a comparison of these two approaches are provided.

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 method with the number of relevant items desired as the stopping criterion is developed. In Section 3.3, the implications of using the number of items desired as the stopping criteria are considered. In the remainder of Section 3, certain important interactions that exist between the definition of precision and the choice of stopping criteria are explained. Then, in Section 4, we provide experimental comparisons between the alternative approaches under consideration and one of the existing methods. Finally, the conclusions of this study are presented in Section 5.

## 2. BACKGROUND

An information retrieval (IR) system is designed and built in response to the need for retrieving useful bibliographic references or texts. Numerous factors contribute to its overall success in satisfying the user population's information requests. In most cases, when a particular search request is presented to a retrieval system, the documents in its collection can be conceptually imagined, to have been divided into two categories. One consists of the set of relevant documents, whereas 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*. Otherwise, it is *nonrelevant*. Hence the usefulness of a retrieval system is determined to a great extent by how closely it can characterize the dichotomy identified above.

In order for a retrieval system to locate the relevant items from a given collection in response to a search request, a measure called the *Retrieval Status Value* (*RSV*) is 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. Many different *similarity* or *distance* functions have been proposed in the past. Among the commonly known examples are the simple matching function and the cosine similarity [25, 32]. None of them has been proved or observed to be optimal under all circumstances. Consequently, the choice of the function for computing the RSVs should be based on the user criterion as expressed by relevant judgments and assumptions about how documents are represented. In any case, 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 RSV ordering 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. However, if more than one item is present at the same level, with an identical RSV, it is termed a weak ordering [6]. Formally, a linear ordering is reflexive, transitive, antisymmetric, and connected (every pair of elements is comparable). In contrast, a weak ordering may not satisfy antisymmetry [29]. 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.

As we stated earlier, numerous system components (for example, the indexing process, internal representation and storage for collection items, the search strategy adopted) affect the ultimate retrieval results and hence the outcome of the performance evaluation. Six different evaluation criteria, deemed most critical to a user population, were pointed out in [9] and [25], 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* 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.

Given a document ranking, 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 if it is assumed that the stopping criterion is the retrieval of $h$ relevant documents, $1 \leq h \leq n$, there are $n$ possible recall levels, that is, $1/n$, $2/n$, ..., $h/n$, ..., $(n - 1)/n$, and 1.

## 2.1 Problem of Weak Ordering

Let $NR$ denote the number of relevant documents that a user desires. 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 desired $NR$ documents. But, when the ordering is not linear, the above method of finding precision must be modified, and some notion of probabilistic precision comes 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 for dealing with this situation was the following: Given $NR$ relevant documents to retrieve (corresponding to a recall level of $NR/n$), we start the search from the very top rank, with the 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 are expected to be retrieved [26, 27, 33, 34]. In other words, the total number of nonrelevant documents that are estimated to be retrieved ($NNR$) is given by

$$NNR = j + \frac{s \cdot i}{r}, \tag{2.1}$$

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

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

We refer to the evaluation method given in eq. (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 3.2.1.

## 2.2 Problem of Multiple Queries

The recall–precision graph is initially defined for a single query. However, in practice, an evaluation result based on a single query is usually not satisfactory. This is because performance comparisons should be made on a sufficiently large number of queries to arrive at statistically significant conclusions. Consequently, assuming an appropriate sample of queries is given, some method of averaging the results from these queries is needed. Since each of the queries might have a different number of relevant documents, the simple recall levels (i.e., $1/n$, $2/n$, ..., $(n - 1)/n$, and 1, previously introduced) cannot 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 $h/n$, $h \geq 2$) is greater than the 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, the precision at each point $k/n$ ($1 \leq k \leq h - 2$) is also 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 the precision at point $h/n$. This whole process is repeated until the last recall point, that is, 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

$$\frac{h}{n} \leq x \leq \frac{h + 1}{n} \quad \text{and} \quad 0 \leq h < n.$$

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, eq. (2.2) becomes

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

The interpolation process above is performed for each query, and the final precision value with respect to each standardized recall is determined by aver-

aging the precision values of all queries at that recall point. Although some other methods of interpolation have been considered in the literature (e.g., [28]), the ceiling method is quite typical of other such methods currently in use.

We refer to *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 demonstrate the problem of interpretation by considering the following examples.

*Example* 2.1. Suppose we have an ordering

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

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 0.25 the precision value estimated by the *PRECALL* method is 0.333.

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

*Example* 2.2. Let the ordering be the 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 0.25 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 0.333: + − −, − + −, − − +. We have

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

where

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

We now obtain

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

since exactly one relevant document is retrieved. For the three arrangements, let $\text{arrangement}_v$ mean that $v$ nonrelevant items are retrieved with that arrangement for getting one relevant item. Then, since $v + 1$ documents are retrieved altogether we get

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

It follows that

$$P(retr \mid \text{arrangement}_0)P(\text{arrangement}_0) = \frac{0+1}{13} \cdot \frac{1}{3} = \frac{1}{13} \cdot \frac{1}{3},$$

$$P(retr \mid \text{arrangement}_1)P(\text{arrangement}_1) = \frac{1+1}{13} \cdot \frac{1}{3} = \frac{2}{13} \cdot \frac{1}{3}$$

and

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

Hence,

$$P(rel \mid retr) = \frac{\frac{1}{13}}{\frac{1}{13} \cdot \frac{1}{3} + \frac{2}{13} \cdot \frac{1}{3} + \frac{3}{13} \cdot \frac{1}{3}} = 0.5.$$

We refer to $P(rel \mid retr)$ 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 is 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, the precision is 1, for the second it is 0.5, and for the third it is 0.333. Hence, for the expected precision, we get

$$EP = 1 \cdot \tfrac{1}{3} + \tfrac{1}{2} \cdot \tfrac{1}{3} + \tfrac{1}{3} \cdot \tfrac{1}{3} = \tfrac{11}{18} \approx 0.611.$$

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 *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* |
|---|---|---|---|
| $\Delta$ | 0.333 | 0.500 | 0.611 |
| $\Delta'$ | 0.375 | 0.444 | 0.609 |

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

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

*Example* 2.5. Let $\Delta = (+ - | + + + + + - - - - - | + + + + - - - -)$, and $\Delta' = (+ + + + + + - - - - | + + - - | + + - - - -)$. For the recall level 0.1, or equivalently for retrieving one relevant document, we obtain precision values for those two definitions as

| | *PRR* | *EP* |
|---|---|---|
| $\Delta$ | 0.667 | 0.750 |
| $\Delta'$ | 0.636 | 0.775 |

From the examples presented in this section we see that when the retrieval output is a weak ordering, there are several ways to define precision for a given simple recall point. Depending 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 a 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 is 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, that is, *PRR* versus *NR*, *PRR* versus *ND*, *EP* versus *NR*, or *EP* versus *ND*. Other stopping criteria are possible; for example, the number of nonrelevant documents that are retrieved (*NNR*) [18]. 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 the 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 $x \cdot n$. Hence, depending on $x$ and $n$, *NR* is not restricted to integer numbers only. For example, if there are 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 are 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 nonrelevant documents searched through in ranks 1 through $(l_f - 1)$.

$r$: number of relevant documents in rank $l_f$.

$i$: number of nonrelevant documents in rank $l_f$.

### 3.2 *PRR* Versus *NR*

3.2.1 *Closed Form Expression for PRR.* In Section 2 we defined $PRR = P(rel \mid retr)$, given that the user requires $NR$ relevant documents. In this section we establish the relationship of $PRR$ to the retrieval system performance measure introduced by Cooper [11], known as the expected search length. Let $P_v$ denote the probability that $v$ nonrelevant documents are retrieved in $l_f$. That is,

$$P_v = P(v \text{ nonrelevant documents retrieved in } l_f \mid s \text{ relevant documents retrieved in } l_f). \tag{3.1}$$

Furthermore, let $s$ denote $NR - t_r$, the number of relevant documents to be retrieved at $l_f$. Then Cooper defines expected search length as the number of nonrelevant documents the user expects to retrieve in an effort to obtain $NR$ relevant items. Notationally we write

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

Using the definition given above we establish the following theorem:

THEOREM 3.1

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

PROOF

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

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

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

If $v$ nonrelevant documents are retrieved in rank $l_f$, then

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

Let $P_v$ be as defined in eq. (3.1). $P(retr)$ can be expressed as a function of the probabilities of certain components by dividing the event "*retr*" into a set of mutually exclusive subevents. For this purpose, let each subevent correspond to the situation that $v$ nonrelevant documents are retrieved in $l_f$. Then

$$\begin{aligned} P(retr) &= \sum_{v=0}^{i} P(retr \mid v \text{ nonrelevant documents retrieved in } l_f) P_v \\ &= \sum_{v=0}^{i} \frac{NR + j + v}{N} \mathrm{P}_v = \frac{NR}{N} + \frac{1}{N} \sum_{v=0}^{i} (j + v) \mathrm{P}_v = \frac{NR}{N} + \frac{1}{N} esl_{NR}. \end{aligned}$$

Hence, we obtain

$$PRR = \frac{NR/N}{NR/N + (1/N) esl_{NR}} = \frac{NR}{NR + esl_{NR}} \qquad \square$$

From Cooper [11] we know that

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

From this we finally obtain

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

Although Cooper derives this expression and interprets it as expected precision, we show here that it is more correctly interpreted as $P(rel \mid retr)$. On the basis of eq. (3.3), Cooper points out that, in computing *esl* versus *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 eq. (2.2). Thus the assumption made about the distribution of documents in $l_f$ is not consistent with that for *PRR* in computing *PRECALL*. This important observation further strengthens our belief that eq. (2.2) may not be used without further justification.

3.2.2 *Multiple Query Evaluation Using PRR.* In Section 2, we established 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 examine two interpolation schemes for *PRR* versus *NR*. The first one is the *ceiling* interpolation similar to what was described in Section 2. The second one, which is called the *intuitive* interpolation, is a new proposal that is more natural than the ceiling interpolation, yet still allows the interpolated values to have meaning as a conditional probability.

When *PRR* versus *NR* is to be calculated under the *ceiling* interpolation, eq. (3.3) instead of eq. (2.3) is used. Specifically, *NR* in eq. (3.3) is replaced by

$\lceil x \cdot n \rceil$ for a given recall level $x$. Except for that change, the interpolation process is identical to what we described in Section 2.2.

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 $x \cdot n$, where $n$ is the total number of relevant documents in response to the given query. We can, therefore, determine *NR* to be 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 eq. (3.3). Hence, we propose the following expression for $0 < s \leq r$:

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

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

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

for integer $s$. If we now interpolate all factorials that contain an $s$ with the $\Gamma$ function we obtain the following lemma.

LEMMA 3.1. *Let esl be calculated by the* $\Gamma$ *function* [16]. *Then,*

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

PROOF. A proof of this lemma is given in the Appendix. The method of proof is similar to Cooper's for integer $s$. □

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 an integer, we can show

$$\sum_{v=0}^{i} P_v = 1, \qquad \text{for all } s.$$

In other words, since the $P_v$'s remain as probabilities, *PRR* continues to have interpretation as a conditional probability for all interpolated values too. Finally, we consider the relationship between the two measures *PRR* and *PRECALL*.

THEOREM 3.2. *PRR versus NR is greater than or equal to PRECALL versus NR.*

PROOF

$$PRECALL = \frac{NR}{NR + j + (s \cdot i)/r} \leq \frac{NR}{NR + j + (s \cdot i)/(r+1)} = PRR \qquad \square$$

In the intuitive interpolation we provide a method for dealing with a 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.

In the first situation we have at least one relevant document in the first rank. From eq. (3.4)

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

Hence $s$ is not involved in the computation of $PRR$.

In the second case there are some ranks that have only nonrelevant documents before the first rank containing relevant documents. Let $j > 0$ be the number of those nonrelevant documents. Again from eq. (3.4)

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

Hence,

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

### 3.3 *PRR* Versus *ND* and *EP* Versus *ND*

Following the ideas of Section 3.2, $PRR$ versus $ND$ is defined as

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

given that the user stops searching after having retrieved $ND$ documents. Let the number of documents to be retrieved in $l_f$ in order to meet the stopping criterion be denoted by $k$. That is, $k = ND - t$. In order to obtain a closed-form formula for $PRR$ we introduce the following lemma:

LEMMA 3.2. *Let $\mu$ be the expected number of relevant documents retrieved in $l_f$. We assume $r$, $i$, and $k$ to be as defined in Section* 3.1. *Then*

$$\mu = \frac{k \cdot r}{r + i}.$$

PROOF. $\mu$ is the expected value of a hypergeometrically distributed random variable [16]. Hence,

$$\mu = \frac{k \cdot r}{r + i}.$$ □

Using the above lemma, we obtain the following theorem:

THEOREM 3.3

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

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

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

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

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

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

$$P(rel \cap retr) = \frac{t_r}{N} + \frac{1}{N} \cdot \frac{k \cdot r}{i + r} \quad \text{by Lemma 3.1.}$$

□

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

$$EP = \sum_{v=0}^{r} \frac{t_r + v}{ND} Q_v.$$

Given the definitions of $PRR$ and $EP$, we now show that $EP = PRR$.

THEOREM 3.4. *If ND is the stopping criterion, then EP = PRR.*

PROOF

$$EP = \sum_{v=0}^{r} \frac{t_r + v}{ND} Q_v = \frac{1/N \sum_{v=0}^{r} (t_r + v) Q_v}{(1/N)ND} = \frac{P(rel \mid retr)}{P(retr)} = PRR$$

□

### 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 the graph in Figure 1, $p$ may *not* be interpreted as either $P(rel \mid retr)$ or as the expected precision corresponding to a recall level of $r$. Thus, explaining the meaning of *PRECALL* is still an open problem. For the examples considered in Section 2.3, these problems remain regardless of the type of interpolation used.

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 is referred to as *intuitive-PRECALL*. Note that in this method eq. (2.2) is applied regardless of whether or not $NR$ is an integer. In order to develop an interpretation for the *intuitive-PRECALL* graph, we define $P(retr \mid rel)$ versus $ND$ and *expected recall* ($ER$) versus $ND$ analogous to the definitions of $PRR$ and $EP$. Then we can prove, in a way similar to that in Theorems 3.3 and 3.4, that

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

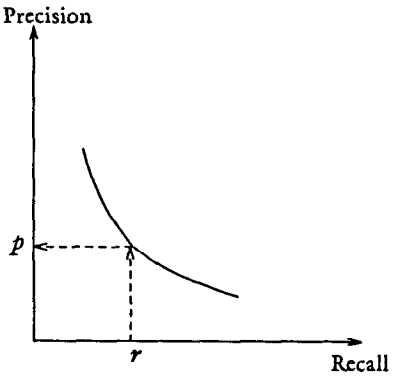


Fig. 1. Interpretation of *PRECALL* as a precision for a given recall.

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

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

for $0 < R \leq 1$. 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 $ND$.

THEOREM 3.5. *Let ER versus ND and EP versus ND be as defined earlier. If* $r \geq 1$, *then* $(ER, EP)$ *with respect to any given integer ND is a point on the graph obtained by the intuitive-PRECALL method.*

PROOF. In order to prove the theorem we have to show that

$$EP = \frac{n \cdot ER}{n \cdot ER + j + (n \cdot ER - t_r)i/r}.$$

This can be seen by substituting

$$ER = \frac{1}{n}\left(t_r + \frac{k \cdot r}{r + i}\right)$$

and

$$EP = \frac{1}{ND}\left(t_r + \frac{k \cdot r}{r + i}\right). \qquad \square$$

We thus obtain the following interpretation of the *PRECALL* Graph: Given any integer $ND$, for $0 < ND \leq N$ and $r \geq 1$, 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 2. This interpretation of the *intuitive-PRECALL* graph requires

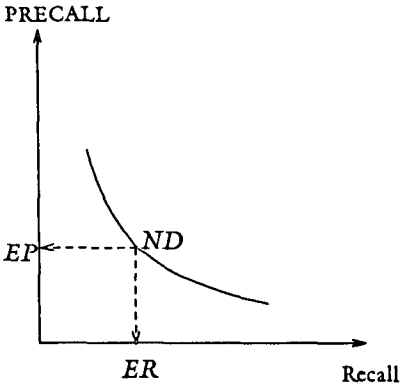


Fig. 2. Parametric interpretation of the graph obtained by the *intuitive-PRECALL* method.

an indirect approach similar to that mentioned in van Rijsbergen [32], where he describes Recall and Precision as a function of a common parameter $\lambda$.

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-à-vis the meaning that can be given to points on the resulting graph. We have not however even been able to find 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 [19] showed that

$$\text{Precision} = \frac{\text{Generality} \times \text{Recall}}{\text{Generality} \times \text{Recall} + (1 - \text{Generality}) \times \text{Fallout}}, \quad (3.5)$$

where Generality $G$ is defined as $G = n/N$ and Fallout is the proportion of nonrelevant documents retrieved. In what follows we want to discuss how the definition of precision as either *PRECALL* or *PRR* is compatible with eq. (3.5). 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 [19]. Hence for any recall $R$ we obtain

$$F = \frac{j}{N - n} + \frac{(n \cdot R - t_r)i}{(N - n)r}. \quad (3.6)$$

If we substitute eq. (3.6) in (3.5) we obtain

$$\frac{G \cdot R}{G \cdot R + (1 - G)F} = \frac{n \cdot R}{n \cdot R + j + (n \cdot R - t_r)(i/r)}.$$

Since $NR = n \cdot R$ and $s = n \cdot R - t_r$ we find out that $GR/(GR + (1 - G)F)$ 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 eq. (3.6). More specifically, given any $(R, F)$ pair, the transformation

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

yields a point on the *intuitive-PRECALL* Graph and vice-versa. Here *PRECALL* may have some meaning indirectly through the (interpolated) $F$ values given by the Recall-Fallout Graph. This depends on whether the proper meaning can be given to the interpolated values of fallout as specified in eq. (3.6). This definition of the Recall-Precision Graph was proposed by Bollmann [1].

Another possibility is to define Fallout for a given recall as the probability of retrieving a nonrelevant document (*PRN*), rather than the expression in eq. (3.6). With this choice we find that

$$\frac{G \cdot R}{G \cdot R + (1 - G)F} = \frac{G \cdot R}{G \cdot R + (1 - G)PRN} = \frac{P(retr \mid rel)P(rel)}{P(retr \mid rel)P(rel) + P(nonrel)PRN}$$

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

Since $F$ is meaningful at noninteger $NR$, the resulting precision is interpretable as a function of recall.

Thus we see that, for a given recall, eq. (3.5) establishes a different notion of precision depending on how fallout is defined. Furthermore, we 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 eq. (3.6) are not interpretable as *PRN* or as the expected value of the ratio of *NNR* to the number of documents retrieved.

## 4. EXPERIMENTAL EVALUATION

On the basis of the two interpolation methods for *PRR* and *PRECALL*, we are interested in exploring the answers to the following. Does *PRR* and *PRECALL* always give us the same conclusions about retrieval performance? Do the two different interpolation methods always give us the same conclusions? What precautions should be taken in utilizing a particular measure? With the *intuitive-PRECALL* method proposed in Section 3.4 what conclusions can be drawn in comparing it to other measures? Specifically, the following three experiments have been carried out to answer such questions.

(1) The first experiment investigated whether by using the same measure, say *PRR*, claims about the relative performance of systems get reversed if we choose different methods of interpolation (i.e., either the ceiling or the intuitive interpolation). In other words, can one measure conclude that retrieval result A is better than retrieval result B, while the other measure leads to the opposite conclusion?

(2) The second experiment investigated retrieval performance comparisons based on two measures: *PRR* under intuitive interpolation and *PRECALL*. To compare the two approaches fairly, we bring them to a common ground by using

the *intuitive-PRECALL* method and by computing the average *PRECALL* over different queries at standardized recall values. Note, however, that although this method is used for making experimental comparisons, its meaning as a function of recall is yet to be determined.

(3) The third experiment examined retrieval evaluation results based on the *intuitive-PRECALL* method where averaging over queries is done at selected *ND* values.

In experiments 1, 2, and 3, four document collections are used. They are ADINUL, CRN4NUL, CISI, and MEDLARS. The collection characteristics are the following:

—ADINUL is a collection of 82 documents in library science. It consists of the full text of papers presented at American Documentation Institute meeting held in 1963. There are 35 queries.

—CRN4NUL has abstracts of 424 documents on aerodynamics, which were used by the Cranfield Project. The corresponding query collection involves 155 queries.

—CISI consists of 1460 documents on library science and has 35 queries. This collection was obtained from Cornell University, where document titles and abstracts were key entered for highly cited information science articles identified by the Institute for Scientific Information.

—MEDLARS is a collection of 1033 documents in the area of biomedicine and has 30 queries associated with it.

To obtain several different retrieval results, we have employed four similarity functions to compute document-query RSVs. The first one is the simple matching function. The second one is the standard cosine similarity. The third one, called "the best probabilistic term weight," and the fourth one, termed "the best (tf × idf)" were proposed in [22] and [23].

For each document collection, retrieval results based on simple matching, cosine similarity, best probabilistic term weight, and best (tf × idf) functions are obtained. Following that, we use the measures *PRR* versus recall based on ceiling interpolation and *PRR* versus recall based on intuitive interpolation to evaluate the impact of the interpolation scheme. Precision values in each of these two measures with respect to recall points 0.1, 0.3, 0.5, 0.7, and 0.9 are then averaged for an overall performance comparison. Tables I, II, III, and IV are the evaluation outcomes respectively for ADINUL, CRN4NUL, CISI, and MEDLARS.

As we can see from these tables, claims do in fact get reversed in some cases. For example, in Table I, *PRR* under ceiling interpolation concludes that the average retrieval result for the cosine function is better than that obtained for the simple matching function. On the other hand, it gives us the opposite conclusion under intuitive interpolation. Other contradictory results occur in the case of the CISI collection when comparing "Best probabilistic term weight" method of "Best tf × idf" method (see Table III). From our discussion in previous sections, we know that *PRR* versus Recall (or *NR*), ignoring the effects of interpolation, is actually not much different from the frequently used measure of *PRECALL* versus *NR* (see Section 3.4). This implies that the validity of conclu-

Table I. ADINUL Collection

| Recall | *PRR* under Ceiling Interpolation | | | | *PRR* under Intuitive Interpolation | | | |
|---|---|---|---|---|---|---|---|---|
| | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf |
| .1 | .3909 | .3787 | .5155 | .5766 | .3522 | .2825 | .4662 | .5598 |
| .3 | .3055 | .3095 | .4163 | .4383 | .3071 | .2703 | .4028 | .4481 |
| .5 | .2188 | .2466 | .3160 | .3960 | .2233 | .2362 | .3293 | .4333 |
| .7 | .1310 | .1357 | .1672 | .2097 | .1326 | .1287 | .1807 | .2390 |
| .9 | .1075 | .1099 | .1307 | .1817 | .1122 | .1126 | .1601 | .2295 |
| Average | .2307 | .2361 | .3091 | .3605 | .2255 | .2061 | .3078 | .3819 |

Table II. CRN4NUL Collection

| Recall | *PRR* under Ceiling Interpolation | | | | *PRR* under Intuitive Interpolation | | | |
|---|---|---|---|---|---|---|---|---|
| | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf |
| .1 | .5306 | .6415 | .7359 | .7103 | .5002 | .5762 | .6847 | .6502 |
| .3 | .3661 | .4515 | .5572 | .5263 | .3664 | .4066 | .5140 | .4812 |
| .5 | .2735 | .3253 | .3964 | .4081 | .2740 | .3021 | .3736 | .3884 |
| .7 | .1637 | .2062 | .2457 | .2526 | .1755 | .1905 | .2297 | .2377 |
| .9 | .0984 | .1249 | .1486 | .1487 | .1059 | .1184 | .1421 | .1421 |
| Average | .2865 | .3499 | .4168 | .4092 | .2844 | .3188 | .3888 | .3799 |

Table III. CISI Collection

| Recall | *PRR* under Ceiling Interpolation | | | | *PRR* under Intuitive Interpolation | | | |
|---|---|---|---|---|---|---|---|---|
| | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf |
| .1 | .2169 | .2483 | .3448 | .3578 | .2270 | .2195 | .3110 | .3144 |
| .3 | .1379 | .1558 | .1825 | .1941 | .1451 | .1521 | .1762 | .1835 |
| .5 | .1011 | .1045 | .1289 | .1401 | .1097 | .1057 | .1335 | .1375 |
| .7 | .0781 | .0791 | .0922 | .0961 | .0889 | .0849 | .1014 | .0958 |
| .9 | .0534 | .0590 | .0634 | .0624 | .0651 | .0674 | .0764 | .0634 |
| Average | .1175 | .1293 | .1624 | .1701 | .1272 | .1259 | .1597 | .1589 |

sions reached by earlier studies using *PRECALL* needs to be questioned and reconsidered owing to the fact that the interpolation technique used previously is unnatural. At least, it is clear that one may not treat the choice of interpolation technique lightly.

In the second set of experiments, we compare the evaluation results given by *PRR* and *PRECALL*. From these experiments we see that the evaluation results

Table IV. MEDLARS Collection

| Recall | *PRR* under Ceiling Interpolation | | | | *PRR* under Intuitive Interpolation | | | |
|---|---|---|---|---|---|---|---|---|
| | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf |
| .1 | .6707 | .7847 | .8561 | .8290 | .6598 | .7287 | .8129 | .7589 |
| .3 | .5134 | .5774 | .6774 | .7066 | .5209 | .5440 | .6419 | .6652 |
| .5 | .3801 | .4347 | .5488 | .5676 | .3896 | .4213 | .5323 | .5530 |
| .7 | .2544 | .3291 | .4019 | .4104 | .2667 | .3222 | .3935 | .3993 |
| .9 | .1255 | .1561 | .2013 | .2113 | .1345 | .1560 | .2001 | .2105 |
| Average | .3888 | .4564 | .5371 | .5450 | .3943 | .4344 | .5161 | .5174 |

Table V. ADINUL Collection

| Recall | *Intuitive-PRECALL* with Averaging by *NR* | | | | *PRR* under Intuitive Interpolation | | | |
|---|---|---|---|---|---|---|---|---|
| | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf |
| .1 | .3302 | .2825 | .4662 | .5598 | .3552 | .2825 | .4662 | .5598 |
| .3 | .2780 | .2703 | .4026 | .4481 | .3071 | .2703 | .4028 | .4481 |
| .5 | .1872 | .2361 | .3290 | .4333 | .2233 | .2362 | .3293 | .4333 |
| .7 | .1193 | .1287 | .1807 | .2390 | .1326 | .1287 | .1807 | .2390 |
| .9 | .0982 | .1125 | .1600 | .2293 | .1122 | .1126 | .1601 | .2295 |
| Average | .2026 | .2060 | .2757 | .3709 | .2255 | .2061 | .3078 | .3819 |

Table VI. CRN4NUL Collection

| Recall | *Intuitive-PRECALL* with Averaging by *NR* | | | | *PRR* under Intuitive Interpolation | | | |
|---|---|---|---|---|---|---|---|---|
| | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf |
| .1 | .4629 | .5762 | .6847 | .6502 | .5002 | .5762 | .6847 | .6502 |
| .3 | .3357 | .4065 | .5138 | .4812 | .3664 | .4066 | .5140 | .4812 |
| .5 | .2471 | .3020 | .3733 | .3883 | .2740 | .3021 | .3736 | .3884 |
| .7 | .1546 | .1903 | .2295 | .2376 | .1755 | .1905 | .2297 | .2378 |
| .9 | .0894 | .1177 | .1412 | .1415 | .1059 | .1184 | .1421 | .1421 |
| Average | .2579 | .3185 | .3885 | .3798 | .2844 | .3188 | .3888 | .3799 |

contradict each other as follows: *PRECALL* contradicts *PRR* for ADINUL and CISI when comparing Simple matching to Cosine (see Tables V and VII).

Finally, we consider the evaluation results obtained by *intuitive-PRECALL* with averaging done over *ND*. These results are summarized in Tables IX–XII. The *ND* values are selected in such a way that the resultant Expected Recall

Table VII. CISI Collection

| Recall | *Intuitive-PRECALL* with Averaging by *NR* | | | | *PRR* under Intuitive Interpolation | | | |
|---|---|---|---|---|---|---|---|---|
| | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf |
| .1 | .2108 | .2195 | .3109 | .3144 | .2270 | .2195 | .3110 | .3144 |
| .3 | .1387 | .1520 | .1761 | .1835 | .1451 | .1521 | .1761 | .1835 |
| .5 | .1059 | .1057 | .1333 | .1374 | .1097 | .1057 | .1335 | .1375 |
| .7 | .0855 | .0848 | .1012 | .0957 | .0889 | .0849 | .1014 | .0958 |
| .9 | .0615 | .0672 | .0761 | .0632 | .0651 | .0674 | .0764 | .0634 |
| Average | .1205 | .1258 | .1595 | .1588 | .1272 | .1259 | .1597 | .1589 |

Table VIII. MEDLARS Collection

| Recall | *Intuitive-PRECALL* with Averaging by *NR* | | | | *PRR* under Intuitive Interpolation | | | |
|---|---|---|---|---|---|---|---|---|
| | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf | Simple Matching | Cosine | Best Probabilistic Term Weight | Best tf × idf |
| .1 | .6412 | .7287 | .8129 | .7589 | .6598 | .7287 | .8129 | .7589 |
| .3 | .5067 | .5435 | .6418 | .6652 | .5209 | .5440 | .6419 | .6652 |
| .5 | .3738 | .4213 | .5320 | .5530 | .3896 | .4213 | .5323 | .5530 |
| .7 | .2546 | .3218 | .3930 | .3992 | .2667 | .3222 | .3935 | .3993 |
| .9 | .1207 | .1547 | .1988 | .2093 | .1345 | .1560 | .2015 | .2105 |
| Average | .3794 | .4340 | .5157 | .5172 | .3943 | .4344 | .5161 | .5174 |

values are distributed evenly in the range [0.0, 1.0]. We get the Expected Recall (*ER*) and Expected Precision (*EP*) for selected *ND* values. For ADINUL the average performance of Cosine in terms of *ER* is better than that of Simple matching, whereas the average *EP* for Simple matching is better than that for Cosine (see Table IX). We see similar *conflicting* results when comparing Simple matching to Cosine for CISI and Best probabilistic term weight to Best tf × idf for both CISI and MEDLARS (see Tables XI and XII).

By comparing the results of averaged *EP* by *intuitive-PRECALL* with averaging done over *ND* with the results of averaged precision by *PRR* (or *intuitive-PRECALL* with averaging done over *NR*), we see that the evaluation results contradict each other as follows:

(1) *intuitive-PRECALL* with averaging done over *ND* contradicts with *intuitive-PRECALL* with averaging done over *NR* for ADINUL and CISI when comparing Simple matching to Cosine (see Tables V, VII, IX, and XI). The same contradiction is also found for CISI and MEDLARS when comparing Best probabilistic term weight to Best tf × idf (see Tables VII, VIII, XI, and XII).

Table IX. ADINUL Collection

| *ND* | Simple Matching | | Cosine | | Best Probabilistic Term Weight | | Best tf × idf | |
|---|---|---|---|---|---|---|---|---|
| | *ER* | *EP* | *ER* | *EP* | *ER* | *EP* | *ER* | *EP* |
| 2 | .1036 | .2646 | .1227 | .2429 | .2002 | .3571 | .2430 | .4000 |
| 10 | .3113 | .1509 | .3573 | .1429 | .4569 | .1800 | .4891 | .2057 |
| 22 | .5077 | .1111 | .5216 | .1078 | .6214 | .1279 | .6522 | .1286 |
| 40 | .7114 | .0854 | .6758 | .0829 | .7963 | .0914 | .8128 | .0936 |
| 68 | .9119 | .0647 | .9173 | .0647 | .9646 | .0681 | .9542 | .0672 |
| Average | .5092 | .1353 | .5189 | .1282 | .6079 | .1649 | .6303 | .1790 |

Table X. CRN4NUL Collection

| *ND* | Simple Matching | | Cosine | | Best Probabilistic Term Weight | | Best tf × idf | |
|---|---|---|---|---|---|---|---|---|
| | *ER* | *EP* | *ER* | *EP* | *ER* | *EP* | *ER* | *EP* |
| 4 | .2123 | .3112 | .2424 | .3484 | .2882 | .4097 | .2933 | .4177 |
| 13 | .4058 | .1879 | .4499 | .2056 | .4995 | .2320 | .4903 | .2318 |
| 38 | .6031 | .0984 | .6690 | .1078 | .6914 | .1123 | .6976 | .1149 |
| 65 | .7076 | .0682 | .7625 | .0728 | .7652 | .0738 | .7869 | .0754 |
| 210 | .9058 | .0276 | .9236 | .0280 | .9117 | .0278 | .9211 | .0280 |
| Average | .5669 | .1387 | .6095 | .1525 | .6312 | .1720 | .6378 | .1736 |

Table XI. CISI Collection

| *ND* | Simple Matching | | Cosine | | Best Probabilistic Term Weight | | Best tf × idf | |
|---|---|---|---|---|---|---|---|---|
| | *ER* | *EP* | *ER* | *EP* | *ER* | *EP* | *ER* | *EP* |
| 15 | .1003 | .2129 | .0866 | .1848 | .1205 | .2686 | .0863 | .2800 |
| 146 | .3664 | .1095 | .3799 | .1173 | .4304 | .1348 | .4315 | .1374 |
| 292 | .5298 | .0840 | .5583 | .0903 | .5988 | .0991 | .6232 | .1015 |
| 584 | .7529 | .0624 | .7685 | .0645 | .7946 | .0670 | .7941 | .0671 |
| 1314 | .9812 | .0372 | .9812 | .0372 | .9812 | .0372 | .9812 | .0372 |
| Average | .5461 | .1012 | .5549 | .0988 | .5851 | .1213 | .5833 | .1246 |

(2) *intuitive-PRECALL* with averaging done over *ND* contradicts *PRR* for CISI and MEDLARS when comparing Best probabilistic term weight to Best tf × idf (see Tables VII, VIII, XI, and XII).

We have drawn the following conclusions from an analysis of the results of these experiments:

(a) One should be aware of the fact that it is possible for evaluation results produced by *PRECALL* and *PRR* to contradict each other.

Table XII. MEDLARS Collection

| *ND* | Simple Matching | | Cosine | | Best Probabilistic Term Weight | | Best tf × idf | |
|---|---|---|---|---|---|---|---|---|
| | *ER* | *EP* | *ER* | *EP* | *ER* | *EP* | *ER* | *EP* |
| 4 | .1204 | .6083 | .1354 | .6667 | .1483 | .7500 | .1370 | .6917 |
| 14 | .3136 | .4738 | .3383 | .5036 | .3886 | .5841 | .3961 | .5976 |
| 35 | .5024 | .3155 | .5447 | .3443 | .6308 | .3935 | .6490 | .4114 |
| 103 | .6986 | .1515 | .7732 | .1717 | .8118 | .1800 | .8297 | .1845 |
| 210 | .9095 | .0512 | .9263 | .0521 | .9341 | .0525 | .9325 | .0525 |
| Average | .5080 | .3201 | .5436 | .3477 | .5827 | .3920 | .5889 | .3875 |

(b) The selection of the method of interpolation is very important since it is possible to reach different conclusions just because of the interpolation technique adopted. Moreover, the proposed intuitive interpolation techniques should be preferred to the so-called ceiling method because of the formal developments given in Section 3.

(c) One should analyze *ER* and *EP* values at the same time for *intuitive-PRECALL* with averaging done over *ND* in order to draw the proper conclusions. Deciding an appropriate set of *ND* values is also a problem.

## 5. 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., the 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* versus *ND* and *PRR* versus *NR* can be handled by relatively efficient computational procedures.

The problem associated with the 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 nonintegral values of *NR* is needed. For *PRR* versus *NR* an interpolation technique that is natural and has a sound formal justification is advanced.

Experiments comparing *PRR* versus *NR* with the method currently well accepted (referred to in this paper as *PRECALL*) are performed and those results are discussed in detail in Section 4. But, as an overall conclusion, we believe that *PRR* versus Recall (or, equivalently, *NR*) has the advantage of having a well-defined meaning. Furthermore, it is closely related to expected search length [11] and lends itself to efficient computation.

In contrast, the *ceiling-PRECALL* method is not amenable to any reasonable interpretation. The problem is caused not only by the fact that averaging results

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 the *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 the *intuitive-PRECALL* method, for averaging purposes, should take precision values over many queries at fixed *ND* (and not *NR*). However, even though intuitive-*PRECALL* gives a sound interpretation, it may have practical difficulties in the selection of *ND*s as follows. When the number of documents in a collection is very large, we must select several *ND*s for which *ER*s and *EP*s are obtained. When *ND* is incremented by fixed intervals, one may not get desired *ER* points that cover a whole range of possible values (i.e., *ER* values may be so close together that one may have difficulty using these as criteria for comparison) since relevant documents are likely to be unevenly distributed among the various ranks. However, with well-selected *ND*s one can use this measure very meaningfully. The question of how to interpret *intuitive-PRECALL* with averaging over *NR* is yet to be addressed. In this paper we also identify the origin of the *intuitive-PRECALL* method and its connection to the Recall-Fallout Graph defined by Robertson [19].

With respect to the other measure *EP*, we show that *EP* versus *ND* coincides with *PRR* versus *ND*. However, the problem of computing *EP* versus *NR* needs to be given a treatment similar to that of *PRR* versus *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 [5].

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 that it helps in the systematic selection of techniques to deal with problems of weak ordering and multiple queries.

## APPENDIX: PROOF FOR LEMMA 3.1

The proof of this lemma is a generalization of Cooper's proof [12] in the sense that we use the $\Gamma$ function here. $\Gamma(x)$ is related to the Beta function [16] by

$$\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \int_0^1 t^{(x-1)}(1-t)^{(y-1)}\,dt = B(x,y) \qquad \text{for} \quad x>0, \quad y>0. \tag{A.1}$$

And we get for $0 < s \le r$

$$\sum_{v=0}^{i}(j+v)\mathrm{P}_v = j + \frac{r!}{\Gamma(s)\Gamma(r-s+1)}\sum_{v=0}^{i} vC_v^i \frac{\Gamma(s+v)\Gamma(i+r-v-s+1)}{\Gamma(r+i+1)}.$$

Putting $s+v$ and $i+r-v-s+1$ to $x$ and $y$ in Equation (A.1), respectively, we obtain

$$\sum_{v=0}^{i}(j+v)\mathrm{P}_v = j + \frac{r!}{\Gamma(s)\Gamma(r-s+1)}\int_0^1 t^{(s-1)}(1-t)^{(r-s)}\left[\sum_{v=0}^{i} vC_v^i t^v (1-t)^{(i-v)}\right]dt.$$

By the binomial theorem we get

$$i \cdot t = \sum_{v=0}^{i} vC_v^i t^v (1-t)^{(i-v)}.$$

Therefore,

$$\sum_{v=0}^{i} (j+v)\mathrm{P}_v = j + \frac{r!i}{\Gamma(s)\Gamma(r-s+1)} \int_0^1 t^s (1-t)^{(r-s)}\, dt.$$

Putting $s+1$ and $r-s+1$ for $x$ and $y$ in Equation (A.1), we have

$$\frac{\Gamma(s+1)\Gamma(r-s+1)}{\Gamma(r+2)} = \int_0^1 t^s (1-t)^{(r-s)}\, dt.$$

Finally, we get

$$\sum_{v=0}^{i} (j+v)\mathrm{P}_v = j + \frac{i \cdot s}{r+1} = esl \qquad \text{for} \quad 0 < s \leq r. \qquad \square$$

ACKNOWLEDGMENT

The authors thank Mr. Shu, Lih-chuyun for providing some of the experimental results used in this paper.

REFERENCES

1. BOLLMANN, P. A comparison of evaluation measures for document retrieval systems. *J. Informatics 1* (1977), 97–116.
2. BOLLMANN, P. Two axioms for evaluation measures in information retrieval. In *Proceedings of the 3d Joint BCS and ACM Symposium.* In *Research and Development in Information Retrieval.* Cambridge, U.K. (1984), pp. 233–245.
3. BOLLMANN, P., AND CHERNIAVSKY, V. S. Measurement-theoretical investigation of the MZ-metric. In *Information Retrieval Research,* R. R. Oddy, et al., Ed., Butterworth, Boston, Mass. 1981.
4. BOLLMANN, P., AND RAGHAVAN, V. V. A utility-theoretic analysis of expected search length. In *Proceedings of the 11th International Conference on Research and Development in Information Retrieval* (Grenoble, France, 1988), Grenoble Univ. Press, pp. 245–256.
5. BOLLMANN, P., RAGHAVAN, V. V., JUNG, G. S., AND SHU, L. Probabiity of relevance and expected precision in evaluating retrieval performance. In preparation.
6. BOOKSTEIN, A., AND COOPER, W. S. A general mathematical model for information retrieval systems. *Libr. Quarterly 46* (1976), 153–157.
7. BUCKLEY, C. Implementation of the SMART information retrieval system. Tech. Rep. 85-686. Dept. of Computer Science, Cornell Univ., Ithaca, N.Y., 1985.
8. CHERNIAVSKY, V. S., AND LAKHUTY, D. G. Problem of evaluating retrieval systems. *I. Naucho-Techniceskaya Informazia,* Ser. 2, pp. 24–30 (in Russian). In English: *Automatic Documentation and Mathematical Linguistics 4* (1970), pp. 9–26.
9. CLEVERDON, C. W. Evaluation of tests of information retrieval systems. *J. Doc. 26* (1970), 55–67.
10. CLEVERDON, C. W. On the inverse relationship of recall and precision. *J. Doc. 28* (1972), 195–201.
11. COOPER, W. S. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *Am. Doc. 19* (1968), 30–41.
12. COOPER, W. S. On selecting a measure of retrieval effectiveness. Part II. Implementation of the philosophy. *J. Am. Soc. Inf. Sci. 24* (Nov./Dec. 1973), 413–424.

13. COOPER, W. S. On selecting a measure of retrieval effectiveness. *J. Am. Soc. Inf. Sci. 24* (1973), 87–100.
14. HEINE, M. H. Distance between sets as an objective measure of retrieval effectiveness. *Inf. Storage and Retrieval 9* (1973), 181–198.
15. HEINE, M. H. The inverse relationship of precision and recall in terms of the Swets model. *J. Doc. 29* (1973), 81–84.
16. HOEL, P. G. *Introduction to Mathematical Statistics*, 4th ed. Wiley, New York, 1971.
17. KRAFT, D. H., AND BOOKSTEIN, A. Evaluation of information retrieval systems: A decision theory approach. *J. Am. Soc. Inf. Sci. 29* (1978), 31–40.
18. KRAFT, D. H., AND LEE, T. Stopping rules and their effect on expected search length. *Inf. Process. Manage. 15* (1979), 47–58.
19. ROBERTSON, S. E. The parametric description of retrieval tests. Part II: Overall measures. *J. Doc. 25* (1969), 93–107.
20. SALTON, G. Evaluation problems in interactive information retrieval. *Inf. Storage and Retrieval 6* (1970), 29–44.
21. SALTON, G., Ed. *The Smart Retrieval System—Experiments in Automatic Document Processing.* Prentice-Hall, Englewood Cliffs, N.J., 1971.
22. SALTON, G. Recent trends in automatic information retrieval. In *Proceedings of the 9th Annual International Conference on Research and Development in Information Retrieval* (Pisa, Italy, Sept. 8–10, 1986). ACM, New York, 1986, pp. 1–10.
23. SALTON, G., AND BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage. 24*, 5 (1988), 513–523.
24. SALTON, G., AND LESK, M. E. Computer evaluation of indexing and text processing. *J. ACM 15*, 1 (Jan. 1968), 8–36.
25. SALTON, G., AND MCGILL, M. J. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.
26. SALTON, G., AND YANG, S. G. On the specification of term values in automatic indexing. *J. Doc. 29*, 4 (1973), 351–372.
27. SALTON, G., YANG, C. S., AND YU, C. T. Contribution to the theory of indexing. *Information Processing 74*, North-Holland, Amsterdam, The Netherlands, 1974, pp. 584–590.
28. SPARCK JONES, K. Performance averaging for recall and precision. *J. Informatics 2* (1978), 95–105.
29. SUPPES, P. *Introduction to Logic.* Van Nostrand, New York, 1957.
30. SWETS, J. A. Effectiveness of information retrieval methods. *Am. Doc. 20* (1969), 72–89.
31. VAN RIJSBERGEN, C. J. Foundations of evaluation. *J. Doc. 30* (1974), 365–373.
32. VAN RIJSBERGEN, C. J. *Information Retrieval,* 2nd Ed. Butterworth Scienti ic Ltd., Surrey, U.K., 1979.
33. YU, C. T., AND SALTON, G. Precision-weighting—An effective automatic i dexing method. *J. ACM 23* (1976), 76–88.
34. YU, C. T., AND RAGHAVAN, V. V. A single-pass method for determining the se nantic relationship between terms. *J. Am. Soc. Inf. Sci. 28* (1977), 345–354.

Received December 1988; revised March 1989; accepted April 1989