CHAPTER 5

EVALUATION OF DOCUMENT RETRIEVAL SYSTEMS

1. The General Problem

An operationally effective automatic document retrieval system must satisfy the requirements of a diverse class of users. The functional model discussed in this thesis considers only a subset of these requirements, principally those related to the methodological aspects of the system operation. The ability to establish system requirements is directly related to the notion of performance evaluation, and in recent years the investigation of evaluation measures for document retrieval systems has recieved considerable attention (references 1 through 8).[*] The purpose of this chapter is to consider in general the problems associated with the design of evaluation experiments and the collection of performance statistics, and to discuss in particular the implications of the system model on these problems.

The general aim of a document retrieval system is to mechanize the deduction of the attribute "relevance", which is a subjective relation between the state of a user's information need and the information content of the "documents" in some collection, as perceived by the user. It is clear that since the system operates

[*] The references cited are only representative of the literature on this topic.

with imperfect representations of both the needs of its users and the information content of documents, and since the notion of relevance is likely to be quite variable over any realistic user population, any performance measure must represent a statistical estimate of the probability distribution of a correct assessment of relevance.

The subjective nature of relevance implies that any realistic system's evaluation will require a large amount of data. Since the collection of such data is costly, and since the notion of evaluation is a critical element of any design process, the designer of a retrieval system must rely on analytical tools, local performance measures, and intuition to select the most likely set of functions to satisfy his objectives. In this connection, computer based simulation systems such as SMART can be of significant benefit since they allow large amounts of data to be generated and analyzed. One of the major objectives of the functional model to be considered here (based on the SMART simulation system) is to maximize the utility of the data generating capabilities of the system by allowing evaluation of the individual functional elements as well as of the overall performance characteristics.

2. Evaluation Measures and the Collection of Statistics

A. The Idealized Experiment

Most of the evaluation measures proposed for document retrieval systems are based on the following idealized characterization of a retrieval operation. A user presents a search request to the system which then compares an index language representation of it to the index images of the documents in some collection. Each comparison results in a binary decision to retrieve or not retrieve the reference document. Independently of the system, it is assumed that the user has made a binary relevance judgment with respect to his information needs (represented by the search request) and the content of each document. The possible results of such an experiment with respect to a single reference document may be represented by the discrete sample space shown in Figure 5.1 (a). Assuming this sample space, estimates of the probabilities associated with each of the sample points (i.e. of the joint probability distribution of the user/system decisions) can be produced by tabulating the number of occurrences of each of the possible outcomes over all of the documents or trials which comprise a single retrieval operation. This is represented by the 2-by-2 contingency table of Figure 5.1 (b), where the ratio of each of the numbers shown to the total number of documents represents the estimate of the probability of the corresponding sample point, i.e.:

$$p_1 = \frac{n_1}{N} = \Pr \left\{\text{Retrieval and Relevance}\right\} , \qquad (5.1)$$

$$p_2 = \frac{n_2}{N} = \Pr \left\{\text{Retrieval and Nonrelevance}\right\} , \qquad (5.2)$$

$$p_3 = \frac{n_3}{N} = \Pr \left\{\text{Nonretrieval and Relevance}\right\} , \qquad (5.3)$$

Sample Point	User Decision	System Decision
1	relevant	retrieve
2	nonrelevant	retrieve
3	relevant	not retrieve
4	nonrelevant	not retrieve

a) The Sample Space of the Outcomes of a Retrieval Operation on a Single Document

	Relevant	Nonrelevant	
Retrieved	n_1	n_2	$n_1 + n_2$
Not Retrieved	n_3	n_4	$n_3 + n_4$
	$n_1 + n_3$	$n_2 + n_4$	$N = n_1 + n_2 + n_3 + n_4$

b) The 2-by-2 Contingency Table of Retrieval and Relevance - the Outcomes of a Retrieval Operation on N Documents

Characterization of Retrieval Results

Figure 5.1

$$p_4 = \frac{n_4}{N} = \Pr\left\{\text{Nonretrieval and Nonrelevance}\right\} . \quad (5.4)$$

From this joint probability distribution a number of conditional probabilities can be defined as follows (following Swets, reference 1):

$$n_1/(n_1+n_3) = \Pr\left\{\text{Retrieval/Relevance}\right\} = p_1/(p_1+p_3), \quad (5.5)$$

the conditional probability of a "hit";

$$n_2/(n_2+n_4) = \Pr\left\{\text{Retrieval/Nonrelevance}\right\} = p_2/(p_2+p_4), \quad (5.6)$$

the conditional probability of a "false drop";

$$n_3/(n_1+n_3) = \Pr\left\{\text{Nonretrieval/Relevance}\right\} = p_3/(p_1+p_3), \quad (5.7)$$

the conditional probability of a "miss";

$$n_4/(n_2+n_4) = \Pr\left\{\text{Nonretrieval/Nonrelevance}\right\} = p_4/(p_2+p_4), \quad (5.8)$$

the conditional probability of a "correct rejection".

Note that $n_1/(n_1+n_3)$ is the recall ratio as defined by Cleverdon[6] while $n_1/(n_1+n_2)$, the conditional probability of relevance given retrieval, is his precision (also called relevance) ratio.

The Bernoulli-like[x] model assumed above is in many respects

[x] A Bernoulli model assumes repeated independent trials in which there are only two possible outcomes for each trial and the probabilities of each outcome remain constant throughout the experiment.

an inadequate description of the true situation. Users do not necessarily make binary relevance decisions nor are such decisions necessarily independent when examining a sequence of documents.[9] In addition, query-document matching functions do not always lead to binary acceptance-rejection decisions; instead, they often result in the assignment of a coefficient of relevance or association between a query-document pair[1] as has been discussed in Chapter 4. Further, in many respects it may be more realistic to assume that the system's assessment of relevance should be interpreted on a relative rather than an absolute basis. Thus, a user is likley to examine at least a few of the highest assessed documents resulting from his search operation, independently of the absolute retrieval coefficients which are assigned to them. In this sense, there is a degree of difficulty in establishing a uniform criterion for what constitutes a positive relevance assessment by the retrieval system over a sample set of search requests.

B. Evaluation Statistics

The contingency table description of a retrieval operation, shown in Figure 5.1 (b) provides frequency ratio estimates of the joint probability distribution of the user/system decisions for the given query. One may then assume, other variables remaining constant, that these frequency ratios converge to probabilities as the number of documents searched (N) increases. Alternatively, one may assume that the probability estimates obtained by a search over N documents predict the behaviour of the system with respect to the input query for some larger collection from which N is a representative random sample.

Consider now the typical evaluation situation in which a number of retrieval operations are performed on some sample set of search requests. In the conceptual framework of the idealized experiment, one assumes the existence of a universe of queries from which the sample set is drawn at random. According to the above characterization, each retrieval operation results in a particular estimate for the joint probability distribution of the user/system decisions, applicable to the input query. Let the query sample contain m elements. The results of the m retrieval operations may be summarized by m 4-tuples:

$$(p_1^i,\ p_2^i,\ p_3^i,\ p_4^i)\quad i = 1,m$$

where the p_k's are defined by equations (5.1) to (5.4). Each of the 4-tuples in addition to defining the probabilities of the sample points of Figure 5.1 (a), defines a set of conditional probabilities such as are given by equations (5.5) to (5.8).

In terms of the probabilistic model, the behaviour of a retrieval system is completely specified by the 4-tuple (p_1,p_2,p_3,p_4) of each query (and associated user relevance decisions) in the universe of queries or query sample space of the system, (which for convience is assumed to be discrete). This sample space defines a joint probability distribution of four random variables P_1, P_2, P_3, and P_4 given by:

$$\Pr\left\{P_1 = p_{1_i}, P_2 = p_{2_j}, P_3 = p_{3_k}, P_4 = p_{4_l}\right\} = f(p_{1_i}, p_{2_j}, p_{3_k}, p_{4_l}) ,$$
$$(i, j, k, l = 1,2,\ldots) .$$

The set of m 4-tuples which result from the test set of retrieval operations provides then an estimate for this joint distribution. The fact that the random variables assume values which are probabilities (or more precisely estimates of probabilities) represents only a notational difficulty.

Statistically then, the objective of an evaluation experiment is to estimate this joint probability distribution or some parameters which characterize it. Clearly any evaluation which ignores the essential fact that the system performance is a random variable defined over the query sample space can produce misleading results. Consider for example the evaluation data produced by the Cranfield studies.[6,7,8] System evaluations in these reports were presented primarily in terms of the two conditional probabilities, precision and recall, rather than in terms of the joint probabilities p_i. This in itself introduces no problem (other than ~~for~~ the fact that it does not represent all the information available in the experimental data); the method used to compute estimates for the mean values of the precision and recall probabilities, however, was in error.

The precision and recall conditional probabilities, being functions of the random variables P_i, are themselves random variables defined on the query sample space. The results of m retrieval operations may be summarized in terms of these conditional probabilities by m couples:

$$(p^i, r^i) \quad i = 1,m \quad ,$$

where

$$\underset{\text{of the i}^{\text{th}}\text{ query}}{\text{precision}} = p^i = \frac{p_1^i}{p_1^i + p_2^i} = \frac{n_1^i}{n_1^i + n_2^i} \; , \tag{5.9}$$

and

$$\underset{\text{of the i}^{\text{th}}\text{ query}}{\text{recall}} = r^i = \frac{p_1^i}{p_1^i + p_3^i} = \frac{n_1^i}{n_1^i + n_3^i} \; , \tag{5.10}$$

where the n_j's are defined by Figure 5.1 (b). The m couples (p^i,r^i) provide an estimate of the joint probability distribution of the random variables P and R defined by:

$$\Pr\left\{P = p_j,\ R = r_k\right\} = g(p_j,\ r_k) \quad (j,\ k = 1,2,..) \ .$$

The respective expectations of these random variables E(P) and E(R) are estimated by the sample means:

$$p = \frac{1}{m} \sum_{i=1}^{m} \frac{p_1^i}{p_1^i + p_2^i} = \frac{1}{m} \sum_{i=1}^{m} \frac{n_1^i}{n_1^i + n_2^i} \quad , \tag{5.11}$$

$$r = \frac{1}{m} \sum_{i=1}^{m} \frac{p_1^i}{p_1^i + p_3^i} = \frac{1}{m} \sum_{i=1}^{m} \frac{n_1^i}{n_1^i + n_3^i} \quad . \tag{5.12}$$

The Cranfield data was interpreted in a different manner. In particular the precision and recall estimates were computed according to the equations:

$$p_c = \frac{\sum_{i=1}^{m} n_1^i}{\sum_{i=1}^{m} n_1^i + n_2^i} \quad , \tag{5.13}$$

$$r_c = \frac{\sum_{i=1}^{m} n_1^i}{\sum_{i=1}^{m} n_1^i + n_3^i} \quad . \tag{5.14}$$

These estimates may be interpreted as resulting from a composite contingency table description of the results of m retrieval operations in which the entries of the composite table are cumulations of the corresponding entries of the m individual tables. As such, these are valid estimates for the conditional population ratios, but not for the means of the associated conditional probabilities over the query sample space.

Without justification it can be assumed that a valid measure of the performance of a retrieval system is the average value received by the system's users. Assuming that the precision and recall conditional probabilities which characterize a given retrieval operation are in fact indicitive of the value of that operation, the estimators defined by equations (5.11) and (5.12) are clearly the appropriate ones. More precisely if it is assumed that the value of the ith of a set of m retrieval operations can be expressed as:

$$v_i = h(p^i, r^i) \quad ;$$

a random variable V is defined which is a function of the random variables P and R. An estimate for the expectation of V, E(V) is given by:

$$v = \frac{1}{m} \sum_{i=1}^{m} h(p^i, r^i) \quad ,$$

i.e. the sample mean, which is a function of the sample distribution of the precision and recall conditional probabilities and not of the population ratio estimates.

A numerical example may serve to illustrate the prece~~e~~ding points. Assume that a sample set of test queries produces results which can be placed in the four categories shown in Table 5.1. It is implied in this hypothetical case that each of the observations is representative of some large subset of input queries of the test sample, so that it can be assumed that the four query types represent equally probable subclasses of the query sample space.

Query Type	n_1	n_2	n_3
	Relevant & Retrieved	Nonrelevant & Retrieved	Relevant & Not Retrieved
1	7	3	3
2	5	5	5
3	9	1	9
4	5	45	45

Retrieval Results for 4 Equally Probable Query Types

Table 5.1

Table 5.2 (a) shows the precision and recall sample distributions and the sample mean estimators for the averages of these random variables over the query sample space. If, however, the data from

Table 5.1 is cumulated into a single set of frequencies, the population ratio estimates for the precision and recall are as shown in Table 5.2 (b). The numbers were chosen to illustrate that the cumulation of observations results in a precision estimate weighted towards the system's performance for queries with higher than average number of documents retrieved, and a recall estimate weighted towards the system's performance for queries with a higher than average number of documents relevant. Thus, for the example shown, the population ratio estimates are biased by the presence of query type 4.

Query Type	Precision	Recall	Cumulative Frequencies		
1	.7	.7	n_1	n_2	n_3
2	.5	.5	26	54	62
3	.9	.5	Population Ratios		
4	.1	.1	$p_c = \frac{26}{80} \approx .33$		
Sample Means	.55	.45	$r_c = \frac{26}{88} \approx .30$		
(a) Query Dependent Statistics			(b) Population Dependent Statistics		

Comparison of Precision and Recall Estimates

Table 5.2

C. Output Characterization

The model discussed above for describing a set of retrieval operations is generally extended by allowing a parametric characterization of search output, i.e. of the system's retrieval decisions. The data obtained in this manner can usually be interpreted as portraying the variation of the joint probability distribution of the user/system decisions as the system's retrieval criterion is relaxed. A typical example of this type of system characterization is given by a precision vs. recall plot such as the one shown in Figure 3.10.

Within the framework of the functional model, the result of a retrieval operation has been characterized by two essentially different forms. Thus as described in Chapter 4, the set inclusion query-document matching function leads to a natural partition of the reference collection into the retrieved and not retrieved subsets. The other matching functions considered (correlation processes) require the specification of a cutoff or decision criterion to induce such a partition. It is shown below that the commom means used to vary the size of the retrieved subset under set inclusion matching is, in fact, equivalent to the use of the set overlap correlation function.

Consider a query containing n_q keywords. With set inclusion matching, the retrieved subset R contains all document images containing at least all n_q query keywords. Define now a subset R(k) which contains all documents that include at least k of the n_q keywords of the query. A uniformly decreasing sequence of values for k from n_q to 1 produces a sequence of retrieved subsets satisfying:

$$R(n_q) \subseteq R(n_q - 1) \subseteq \ldots . \subseteq R(2) \subseteq R(1) \quad .$$

The retrieved subset R(k) is thus monotonically increasing with decreasing values of the cutoff parameter k.

The set overlap correlation function was defined in Chapter 4 as:

$$\rho(q,d) = n(q \cap d),$$

where n(A) is the number of elements in the set A. Thus if the query q contains n_q keywords, all documents containing at least those same n_q keywords receive correlation n_q. Documents containing n_q-1 of the n_q query terms receive correlation n_q-1, etc. Therefore, the union of all document subsets with correlation k or greater under the overlap matching function is equivalent to the retrieved set R(k) under set inclusion matching, and thus when the retrieval criterion is allowed to vary, these matching functions are essentially equivalent.

Set represented index images lead to retrieval rankings of document subsets, whereas with vector represented index images individual documents are ranked. The difference is essentially one of degree and can be attributed to the increased information content of the vector index language. With a vector correlation matching process the retrieved subset may be parametrically associated with a cutoff correlation (defined either absolutely or relatively with respect to the correlation distribution for each query), or with the rank position of documents in the ordering induced from the correlation coefficient (for example by defining the retrieved subset to contain the k highest correlating documents). The common property of any of these alternatives is that they all yield a sequence of monotone increasing retrieved subsets as the degree of association (defined by the matching relation) is decreased. Each of these techniques is commonly used to represent the variation in the joint distribution of the user/system decisions as the quantity of output (the size of the retrieved subset) is increased (or equivalently as the matching criterion is relaxed). In section 4 of this chapter an alternative to the general evaluation strategy of describing performance by a set of parameters which vary with discrete changes in the matching criterion is presented.

D. The Precision-Recall Tradeoff

The use of a precision vs. recall plot variable with the cutoff parameter as an evaluation tool for document retrieval systems (introduced by Cleverdon[6]) has led to observations that there exists a so-called tradeoff between these two conditional probabilities which is of fundamental significance. It will be shown here, however, that this inverse relationship is a direct consequence of assuming a statistically significant matching function, and further that both of these conditional probabilities are increased by any process which improves the joint probability of retrieval and relevance.

The increase in recall as the amount of output accepted as retrieved is increased is a direct consequence of the definition of the recall conditional probability. Since the retrieved subset is monotonically increasing, the ratio of relevant documents retrieved to total number of relevant documents (a constant for any retrieval operation) is necessarily monotonically increasing. Precision, however, is defined as the ratio of relevant documents retrieved to the total number of retrieved documents. It has been shown above that an increase in the size of the retrieved subset is tantamount to relaxing the requirements for query-document matching. Thus if it is assumed that the matching function is a statistically significant indicator of relevance (the counter assumption is clearly contradictory to its use), precision must decrease with increase in the size of the retrieved subset.

As a concrete example, consider the vector indexing model. With respect to a given query, one assumes that the probability that a document $\bar{d}_i$ be relevant to the query is a monotonic function of the correlation coefficient $\rho(\bar{q},\bar{d}_i)$. Consider the two highest correlating documents $\bar{d}_1^i$ and $\bar{d}_2^i$ which result from search operations for some ensemble of queries $\bar{q}_i$. Let q_1 be the probability that $\bar{d}_1$ is relevant and q_2 be the probability that $\bar{d}_2$ is relevant. The assumption above implies that averages over the query ensemble will yield estimates for these probabilities such that:

$$\hat{q}_1 > \hat{q}_2 \ . \tag{5.15}$$

Now assume that the precision ratio is calculated after each retrieved document (i.e. the cutoff is a function of the retrieval ordering). At cutoff 1, the precision ratio is clearly $\hat{q}_1$. At cutoff 2, the precision ratio is $(\hat{q}_1+\hat{q}_2)/2$. Since $\hat{q}_1 > \hat{q}_2$ implies that $\hat{q}_1 > (\hat{q}_1+\hat{q}_2)/2$, the precision decreases as the number of documents considered retrieved, increases.

The tradeoff, then, between precision and recall is a necessary statistical consequence of using a meaningful matching function. The nature of this tradeoff is fundamentally related to the joint probability distribution of the user/system decisions from which the conditional probabilities, recall and precision, are defined. Improvements in retrieval systems which increase the joint probability of relevance and retrieval will increase both recall and precision for a given level of query-document association. For a given user, the inverse relation of recall and precision influence the number of output (retrieved) documents which it is useful for him to examine.[2]

3. The Use of Optimal Queries in Test Design

In Chapter 3 the notion of an optimal search request was introduced and developed from the point of view of query modification in a system environment allowing iterative searches, and real time system-user interaction. It was noted there that the concept of an optimal query offered the potential of allowing an explicit evaluation of the power of the index language independent of the performance variations which can be expected from the query formulation process.

In essence, any evaluation measure based on a retrieval operation with an optimal query is a measure of the relative association between the members of a subset of relevant documents (specified by the user) compared to the association of these documents with the entire collection. Viewed in this manner, the definition of an optimal query offers a positive alternative to the design of evaluation experiments for retrieval systems. Conventionally, on an experimental basis, there is a serious problem in obtaining representative search requests with accompanying relevance judgments (witnessed by the controversy as to the value of the test queries used in the Cranfield study[10] for example). Rather than measuring performance by the use of test queries, a retrieval system can be evaluated by obtaining user judgments as to the degree of association between documents which by their location in the index space are necessarily associated.

Such an evaluation represents a measure of the ability of the indexing scheme to preserve the associations in the index space which users can detect from the information content as expressed in the natural language. Since the retrieval performance which a user can expect is a function of the degree to which input search requests correspond (or can be adjusted to correspond) to their respective optimal forms, and since the performance of an optimal query is directly related to the consistency of the associations of documents in the index space, a measure of the latter is indicative of a measure of the former. While such a procedure does not mecessarily reduce the quantity of subjective data required for a significant test of a retrieval system, it offers the potential for providing an increased measure of experimental control. Since the requirements for examining the implications of such a test program are prohibitive with respect to the scope of this thesis (by virtue of requiring a large subjective test effort), this test design is offered as a suggestion worthy of additional consideration.

4. Cutoff-Independent Performance Indices

A. Derivation

Performance indices for document retrieval systems which are based on a contingency table description (as introduced in section 2) assume that a retrieval operation partitions the reference collection, i.e. identifies a retrieved subset. In the model system considered in this thesis (vector indexing and cosine correlation query-document matching), and in other models of interest (see Chapter 4), the result of a retrieval operation is more accurately described by the distribution of the matching coefficient over the reference collection or by the ordering induced on the document set from this distribution. The use of partition based evaluation parameters for such systems requires, then, that some decision function (or cutoff criterion) be introduced into the retrieval process. Operationally, the number of retrieved documents a user will examine is likely to be dependent on a number of subjective variables. There is, therefore, considerable difficulty in the a priori specification of a meaningful partitioning algorithm. For this reason, then, some performance measures are derived here which are functionally dependent on the full ordering of the reference collection produced by a retrieval operation. Such measures eliminate the need to introduce any notion of cutoff.

Under the assumption that the ordering induced on the set of reference documents by the search process M is the principal result of a retrieval operation and that a set of relevant documents D_R is available corresponding to each request q, the objective of a retrieval operation may be expressed as follows: a retrieval operation with respect to a request q is expected to produce an ordering on the reference collection D, such that every member of the set D_R is ranked above all members of the complement of D_R with respect to $D(\bar{D}_R)$.

Note that in this formulation no emphasis is placed on any relative order among the members of the set D_R of relevant documents. While such an ordering might in theory seem desirable, the determination of an unordered set D_R is difficult enough by itself, so that imposition of an additional ordering criterion may be impractical. A partial order within D_R may, however, have some significance and, in fact, has been employed in some of the ASLIB-Cranfield experiments to specify degrees of relevance. These in turn lead to the definition of different subsets D_R, but not to the specification of retrieval order with respect to relevance order.

Given the previously stated definition of the objective of a retrieval operation, two functions of the ordering induced on D may be defined which are related to the recall and relevance (precision) of Cleverdon. Consider an ordering induced on D by M such that a one-to-one mapping exists from D to the dense set of integers from 1 to n(D); increasing rank order in the set of integers then reflects decreasing connection between the request image and document image.

In this case, define:

$$r^*(i) = \begin{cases} \frac{i}{n_o} \text{ for } 1 \leq i \leq n_o \\ 1 \text{ for } n_o \leq i \leq N \end{cases}$$

and

$$p^*(i) = \begin{cases} 1 \text{ for } 1 \leq i \leq n_o \\ \frac{n_o}{i} \text{ for } n_o \leq i \leq N \end{cases},$$

where

$n_o = n(D_R)$, i.e., the number of relevant documents to the query under consideration;

$N = n(D)$, the number of documents in the reference collection; and

i = the rank index induced on D.

The function r(i) is viewed as the number of relevant documents having rank order less than or equal to i divided by the total number of relevant documents. Thus, it is Cleverdon's recall as a function of the order induced on D by a retrieval operation. Clearly, r*(i) is the recall function which pertains when the retrieval operation produces an ideal ordering on D. Similarly, p(i) is the number of relevant documents having rank order less than or equal to i divided by i, with p*(i) defined for the case when all members of D_R have a rank index less than every member of $\bar{D}_q$. Hence for each query q, $r^*_q(i)$ defines a desired (or objective) recall function, and $p^*_q(i)$ defines a desired precision function.

Since it has been assumed that M induces only an ordering on D, as opposed to a metric, these functions are strictly defined only for discrete values of the rank index i. As it is intended to extend these functions to a continuous independent variable, that is, to define a function $r^*(x)$ equal to $r^*(i)$ for $x = i$, a possible anomaly is noted. This arises from the fact that it is possible, within the framework of the system, for M to produce a mapping from elements of $\{q \times D\}$ to the real line. This, in fact, occurs when M is a correlation process which correlates a query image with the set of document images viewed as vectors in some abstract space. The process of inducing an ordering from this mapping and then treating this ordering as a function of a continuous real variable gives the impression of coming full circle. In fact, there is clearly a loss of information involved since relative distance between the images of $\bar{d}_i$ and $\bar{d}_j$ is not preserved by this process. The justification for making this transformation from the domain of M to an ordering index lies in the assumption that the order so derived has significance of and by itself.

The extension then to functions of a real variable is accomplished by defining two functions $r^*(x)$ and $p^*(x)$ such that:

$$\left.\begin{aligned} r^*(x) &= r^*(i) \\ p^*(x) &= p^*(i) \end{aligned}\right\} \text{ for } x = i,\ i = 1,2,\ldots,N \ ;$$

and further that:

$$r^*(x) = \begin{cases} \dfrac{j}{n_o} \text{ for } j \leq x \leq j+1 \text{ , and } j \text{ integral and less than } n_o \\ 1 \text{ for } x > n_o \end{cases} ;$$

and

$$p^*(x) = \begin{cases} \dfrac{j}{x} \text{ for } j \leq x < j+1\text{, and } j \text{ integral and less than } n_o \\ \dfrac{n_o}{x} \text{ for } x > n_o \end{cases} .$$

At this point, recall and precision functions may be defined for the results of a retrieval operation with respect to a particular query. In particular, let the ranks of each member of the set of relevant documents D_R resulting from applying M to $\{q \times D\}$ be specified as:

$$\emptyset(i) \quad \text{for } i = 1,2,\ldots,n_o,$$

where $\emptyset(i + 1) > \emptyset(i)$.

In this case:

$$r_q(x) = \begin{cases} 0 & \text{for } 1 \leq x \leq \emptyset(1) \\ \frac{i}{n_o} & \text{for } \emptyset(i) \leq x < \emptyset(i+1) \\ 1 & \text{for } x \geq 0(n_o) \end{cases} \quad ;$$

and

$$p_q(x) = \begin{cases} 0 & \text{for } 1 \leq x \leq \emptyset(1) \\ \frac{i}{x} & \text{for } \emptyset(i) \leq x < \emptyset(i+1) \\ \frac{n_o}{x} & \text{for } x \geq \emptyset(n_o) \end{cases} \quad .$$

At this point recall and precision error functions may be defined by the equations:

$$\text{recall error} = \int_{x=1}^{N} (r^*(x) - r_q(x))dx \quad ; \qquad (5.16)$$

$$\text{precision error} = \int_{x=1}^{N} (p^*(x) - p_q(x))dx \quad . \qquad (5.17)$$

Since r*(x) is an upper bound to r(x) and p*(x) is an upper bound to p(x), the error functions are always greater or equal to zero.

To evaluate these integrals the unit step function, $U_{-1}(x)$ is introduced, defined by:

$$U_{-1}(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \quad ;$$

for which:

$$\int_{-\infty}^{b} U_{-1}(x)dx = b$$

Now r*(x) can be expressed as:

$$r^*(x) = \frac{1}{n_o}\left[U_{-1}(x-1) + U_{-1}(x-2) + \ldots . + U_{-1}(x-n_o)\right] \quad ,$$

and

$$r(x) = \frac{1}{n_o}\left[U_{-1}(x-\varnothing(1)) + U_{-1}(x-\varnothing(2) + \ldots . + U_{-1}(x-\varnothing(n_o)\right] \quad .$$

Therefore,

$$\int_{1}^{N} (r^*(x) - r(x))dx = \frac{1}{n_o} \sum_{i=1}^{n_o} \int_{1}^{N} \left[U_{-1}(x-i) - U_{-1}(x-\varnothing(i)\right] dx$$

$$= \frac{1}{n_o} \sum_{i=1}^{n_o} \left[\varnothing(i) - i\right]$$

$$= \frac{1}{n_o} \sum_{i=1}^{n_o} \varnothing(i) - \frac{1}{n_o} \sum_{i=1}^{n_o} i$$

or

$$\text{recall error} = \bar{\varnothing} - \frac{n_o + 1}{2} \quad . \tag{5.18}$$

Thus the integral of the difference between the recall function for a perfect retrieval and the recall function of an actual retrieval operation is the difference between the actual average rank ($\bar{\varnothing}$) of the members of the set of relevant documents D_R, and the average rank $(n_o+1)/2$ which would obtain under perfect retrieval.

This parameter may be normalized to the range 0 - 1 by considering the case for which the rank of every member of D_R is numerically greater than every member of $\bar{D}_R$. This is clearly the case of maximum error; therefore:

$$\text{max recall error} = \frac{1}{n_o} \sum_{i=1}^{n_o} N - (i-1) \quad - \frac{n_o + 1}{2}$$

$$= \frac{1}{n_o}\left[\frac{n_o}{2}(N + N - n_o + 1)\right] - \frac{n_o + 1}{2}$$

$$= N - n_o \quad .$$

Hence:

$$re_n \quad = \quad \frac{\bar{\varnothing} - \frac{n_o + 1}{2}}{N - n_o} \tag{5.19}$$

is a normalized index of the recall error. As this parameter measures recall error, it is desirable to reverse it. Therefore:

$$r_n \quad = \quad 1 - \left[\frac{\bar{\varnothing} - (\frac{n_o + 1}{2})}{N - n_o}\right] \tag{5.20}$$

is the desired normalized performance index.

The precision error integral may be evaluated through the use of the unit step function as follows:

$$p^*(x) = \frac{1}{x}\left[U_{-1}(x-1) + U_{-1}(x-2) + \ldots. + U_{-1}(x-n_o)\right] ,$$

and

$$p(x) = \frac{1}{x}\left[U_{-1}(x-\emptyset(1)) + U_{-1}(x-\emptyset(2)) + \ldots. + U_{-1}(x-\emptyset(n_o))\right] .$$

Now

$$\int_{-\infty}^{b} U_{-1}(x-a)\,\frac{dx}{x} = \int_{a}^{b} \frac{dx}{x} = \ln b - \ln a .$$

Therefore:

$$\text{precision error} = \int_{1}^{N} (p^*(x) - p(x))dx$$

$$= \sum_{i=1}^{n_o} \int_{1}^{N} \frac{dx}{x}\left[U_{-1}(x-i) - U_{-1}(x-\emptyset(i))\right]$$

$$= \sum_{i=1}^{n_o} \ln \emptyset(i) - \sum_{i=1}^{n_o} \ln i ,$$

or

$$\text{precision error} = \ln \prod_{i=1}^{n_o} \emptyset(i) - \ln (n_o!) . \quad (5.21)$$

This index may be normalized to lie in the range 0 - 1 by dividing by the maximum possible error. From the previous argument this error is:

$$\text{max precision error} = \ln \prod_{i=1}^{n_o} N - (i+1) - \ln (n_o!)$$

$$= \ln \frac{N!}{(N - n_o)!} - \ln(n_o!)$$

$$= \ln\binom{N}{n_o} \text{, (the binomial coefficient).}$$

The normalized index of precision error is therefore:

$$pe_n = \frac{\ln \prod_{i=1}^{n_o} \not{0}(i) - \ln(n_o!)}{\ln\binom{N}{n_o}} . \qquad (5.22)$$

Again, since this is an index of precision error, it is desirable to reverse it. Therefore:

$$p_n = 1 - \frac{\ln \prod_{i=1}^{n_o} \not{0}(i) - \ln(n_o!)}{\ln\binom{N}{n_o}} \qquad (5.23)$$

is the desired normalized index of precision performance.

Since both these indic~~i~~es reflect over-all system performance a value of 1 for either implies a value of 1 for the other, in contrast to the conventional recall and precision evaluation measures. The difference between these two over-all measures lies in the weighting given to the relative position of the relevant documents in the ordered retrieval list. The recall index (equation (5.20)) weights rank order uniformly, and is therefore equally sensitive to the rank of every relevant document. The precision index (equation (5.23)), however, weights initial ranks more strongly, and is therefore more sensitive to the system's behavio~~u~~r as reflected by the initial distribution of retrieved documents.

The recall and precision indices derived here depend on the assumption that the ordering induced on D by M is a full order, i.e., that it can be represented by a one-to-one mapping from D to the dense set of integers from 1 to n(D). In general this may not be the case since a partial order rather than a full order may result from a given retrieval operation; therefore a method for defining document rank in this event is required.

The most natural way of treating documents which are equivalent under a partial retrieval ordering is to give each member of the equivalent set the average of the ranks which would apply to the set members if they were differentiable. Hence, if M induces the partial order: $d_1 > d_2 > \{d_3, d_4, d_5\} > d_6$ on a set $D = \{d_1, d_2, d_3, d_4, d_5, d_6\}$, ranks are assigned in the sequence: 1,2,4,4,4,6.

In the derivation above of the normalized rank recall (eq. (5.20)) and the normalized log precision (eq. (5.23)) it was assumed that all members of the set of relevant documents D_R were of equal value. Consider now an extension of these indices by assuming that a partial ordering on D_R is specified which reflects degree of relevance, i.e.,

$$D_{R_1} > D_{R_2} > \dots > D_{R_k} \quad ,$$

where $D_{R_i} \in D_R$ and $>$ implies "more relevant than." In this case the objective of a retrieval operation may be defined as follows: a retrieval operation with respect to a query q and a partially ordered set of relevant documents D_R is expected to produce an ordering on the reference collection D, such that every member of the set D_{R_i} is ranked before all members of the sets D_{R_k} for which $i < k$, and that all members of D_R are ranked before members of $\bar{D}_R$. Corresponding to this definition, expressions for $r^*(x)$, $p^*(x)$, $r_q(x)$, and $p_q(x)$ may be defined in a manner analogous to those previously used. The development of the performance indices for this case is more cumbersome than for the case presented above. As the situation to which these extended indices are applicable is not normally considered to be of general interest their derivation is omitted.

B. Experimental Use

The performance measures developed in this section have been used to evaluate the results of a variety of experiments conducted with the SMART system.[11,12] As one might expect from the formulation, the range of the normalized recall index is rather limited; i.e. a random retrieval yields an expected recall index of .5, hence one would suspect results observed in practice to be close to 1.0. In fact, the observed range of this index from a variety of SMART system experiments is from about .9 to 1.0, with an average near .97. The normalized precision index however, being more dependent on the initial part of the retrieved sequence, exibits a reasonable range for the search requests examined to date, and typically varies from .6 to 1.0. In practice then, to produce a useful range of values for the recall index, one is forced to expand its scale. A scale expansion of 5, introduced so as to maintain an upper bound of 1.0, produces an observed range for the scaled recall index similar to that of the precision index. The scaled index is defined as:

$$r_{n_s} = 1.0 - 5(1.0 - r_n)$$

where r_n is the normalized rank recall defined by equation (5.20).

Two related performance indices may be derived from the two which have been considered. These are useful in the case where a particular query is subjected to a set of retrieval operations (varying some system parameter for example) which are to be compared. The recall error, equation (5.18) was derived as:

$$\text{recall error} = \bar{\emptyset} - \frac{n_o + 1}{2} \quad .$$

Since $\bar{O}_{min} = (n_o+1) / 2$, a positive index with an upper bound of 1.0 may be defined as:

$$\text{rank recall} = \frac{\dfrac{n_o + 1}{2}}{\bar{\emptyset}} \quad .$$

A similar observation for the case of the derived precision error, equation (5.21), produces the index:

$$\text{log precision} = \frac{\ln (n_o!)}{\ln \prod_{i=1}^{n_o} \phi(i)}$$

The advantage of these indices lies in the fact that they are simpler and therefore easier to compute than the normalized indices. In addition, the rank recall parameter requires no scale expansion to assume a useful range. The disadvantage of both these measures is their dependence on n_o, the number of relevant documents. As n_o varies from query to query this dependence makes it impossible to average the unnormalized indices over a set of search requests to produce a meaningful system's evaluation.

REFERENCES

1. Swets, J.A., "Information Retrieval Systems", Science, Vol. 141, July 1963
2. Verhoeff, J., Goffman, W., and Belzer, J., "Inefficiency of the Use of Boolean Functions for Information Retrieval Systems", Communications of the ACM, Vol. 4, No. 12, Dec. 1961
3. Swanson, D.R., "Searching Natural Language Text by Computer", Science, Vol. 132, Oct. 1960
4. Swanson, D.R., "Interrogating a Computer in Natural Language", Information Processing 1962, Proceedings of IFIP Congress 62, Amsterdam North-Holland Publishing Company, 1963
5. Swanson, D.R., "Design of Experiments for the Testing of Indexing Systems", draft manuscript
6. Cleverdon, C.W., "Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems", ASLIB - Cranfield Research Report, Cranfield, Oct. 1962
7. Cleverdon, C.W., "The Testing of Index Language Devices", ASLIB Proceedings, Vol. 15, No. 4, April 1963
8. Lancaster, F.W. and Mills, J., "Testing Indexes and Index Language Devices: ASLIB Cranfield Project", American Documentation, Vol. 15, No. 1, Jan. 1964
9. Goffman, W., "A Searching Procedure for Information Retrieval", Information Storage and Retrieval, Vol. 2, No. 1, Jan. 1964
10. O'Connor, J., "A Review of the Cranfield Project", Journal of Documentation, Vol. 17, 1961

11. Salton, G., "The Evaluation of Automatic Retrieval Procedures - Selected Test Results Using the SMART System", American Documentation, Vol. 16, No. 3, July 1965

12. Rocchio, J., and Engel, M., "Test Design and Detailed Retrieval Results", Report ISR 8, The National Science Foundation, Harvard Computation Laboratory, Dec. 1964