Evaluating the Small Information Retrieval System*

F. W. LANCASTER
National Library of Medicine,
8600 Wisconsin Ave., Bethesda, Maryland 20014

Received June 21, 1966

Evaluation is an analytical procedure: We analyze to determine how far the system is satisfying user requirements, we analyze to determine sources of system failure, and we analyze to determine how these failures may best be remedied.

This paper deals less with details of methodology than with reasons for evaluating an operating retrieval system, and with the possible benefits that can accrue from a properly conducted test program. These benefits may take the form of improved service to users, or reduction in costs of service, of possibly both. Viewed in this light, an evaluation program can be important to the small information system as well as to the large. In fact, because of its more limited resources, and the consequent need to exploit these to the full, systems evaluation may be of even greater value to the smaller information group.

Despite a considerable body of literature to the contrary, the prime requirements of users of an information retrieval system are self-evident. The tolerances of users in relation to these requirements, and in relation to the tradeoffs between them, will vary from situation to situation, but the major requirements remain the same. The prime concern is the ability of the system to retrieve documents of some value to the user in relation to the information need that prompted his request. In other words, he is interested in the *recall* power of the system. In addition, he is concerned with the ability of the system to screen out and hold back documents of no value to him. That is, he is interested in the filtering capacity of the system, or its *precision* capabilities. Precision is one measure of the effort required—by system user, system operator, or both—in order to obtain a particular recall figure. To measure recall without considering precision, and vice versa, is meaningless. Obviously, we can always obtain maximum recall for any request by retrieving the entire collection. In such a case, the filtering mechanism of the system—*i.e.*, the index—has not been brought into play at all. To measure the ability of the system to let through wanted documents, but to hold back unwanted ones, we must consider recall ånd precision jointly. Although users are concerned with other things besides recall and precision (for example, response time, form of presentation of results, and the amount of effort they must expend in order to consult the system), these other factors are strictly secondary or are factors influencing recall and precision tolerances.

As Fairthorne (*1*) has pointed out, the results of a search, for any request, can be expressed as components of a 2 x 2 contingency table, as follows:

	Wanted Documents	Unwanted Documents
Retrieved	Wanted and retrieved	Unwanted but retrieved
Not Retrieved	Wanted but not retrieved	Unwanted and not retrieved

When we undertake a search under controlled conditions within the context of an evaluation program, we must be able to derive the appropriate figures and insert them into this 2 × 2 table. While there are a number of ways in which the results of a search, or series of searches, can be presented, one convenient method that has proved useful in previous studies is to use *recall ratios* and *precision ratios*.

$$\text{Recall ratio} = \frac{\text{number of wanted documents retrieved}}{\text{total number of wanted documents in the collection}}$$

while

$$\text{Precision ratio} = \frac{\text{number of wanted documents retrieved}}{\text{total number of documents retrieved}}$$

Assume a request for which we establish, by some analysis procedure, that there are 10 relevant documents in the collection. If we retrieve eight of these in searching, we say that our recall ratio for this search is $^8/_{10}$, or 80%. At the same time, if our total retrieval has been 100 citations, or documents, our precision ratio for the search is $^8/_{100}$ (*i.e.*, eight wanted, 92 unwanted), or 8%.

Of course, we cannot evaluate system performance on the basis of its response to a single request. We must evaluate performance in relation to a significant number of representative requests. These test requests should pref-

* Presented before the Division of Chemical Literature, Symposium on Problems of Small Information Groups, 151st National Meeting of the American Chemical Society, Pittsburgh, Pa., March 25, 1966.

erably be real-life requests, representing actual information needs, made to the retrieval system over a particular time period. One way to obtain a representative sample would be by establishing a simple sampling rate—for example, taking every fifth request over a six-month period. The performance figures for the system are established by averaging the performance figures obtained for the individual requests.

The establishment of recall ratios and precision ratios requires that decisions be made, in relation to a particular request, as to which documents are "wanted" and which are "unwanted." Although the question of relevance generates considerable controversy, in the evaluation of an operating retrieval system, where the entire system is being evaluated, a "relevant" or "wanted" document is nothing more nor less than a document of some value to the requester in relation to the information need that prompted his request. Obviously, then, relevance decisions are value judgments on documents made by a requester in relation to his information need.

It is extremely important, for the subsequent analysis, that the requester be asked to substantiate his value judgments by indicating why certain documents are, in relation to his information need, of major value, others of minor value, and others of no value.

When a requester evaluates the documents retrieved in a search, or a random sample of them, and indicates which are wanted and which unwanted, he has made the decisions that will allow us to calculate a precision ratio for this particular search. Establishment of a recall ratio for a search is rather more difficult. The ideal of having the requester examine the entire collection, while feasible in certain experimental conditions, is obviously not feasible in the evaluation of an operating system. Conventional random sampling of the residual file—*i.e.*, the file of documents *not* retrieved by the search—is usually impractical because of the size of the sample that would need to be drawn and examined in order to have any expectation of finding even one relevant document.

There appear to be essentially only two possible ways of establishing a recall figure for an operating retrieval system. The first involves the use of prepared requests (requests formulated on the basis of documents known to be in the files, that is, *source documents*). Recall ratios and precision ratios can be derived for a group of prepared requests, and can be matched against precision ratios obtained for a group of real-life requests, thereby allowing derivation of recall ratios for the real requests by a technique of extrapolation.

Source documents have been used, with success, to establish a recall figure in a number of system evaluations (*2, 3*). Although the method has been criticized, it can yield, with careful controls, an estimate of recall that will be adequate for many purposes of evaluation. Because it is simple and economical to implement, this method is particularly applicable in the evaluation of the small retrieval system.

However, the advantage of being able to derive recall and precision figures for a group of real requests, without the need to use prepared requests, is undeniable. The method that shows greatest promise is that currently being used in an evaluation of MEDLARS at the National Library of Medicine. Recall ratios are established on the basis of system performance in relation to a group of documents *found by means outside the system being evaluated*, but confirmed to be in the data base of the system, and judged by the requester to be relevant to the request. Suppose, for a particular request, we discover 20 documents by means extraneous to the system—*e.g.*, items known by the requester in advance of search, items suggested by authors of papers cited by the requester, items found by a search in some specialized bibliography or information center. Let us also suppose that 18 of these items are found to have been indexed into our system, and, of these 18, the requester judges 15 to be relevant. If in our search, we retrieve 12 of these 15 items, we can say that our recall ratio for this request is $^{12}/_{15}$ or 80%.

It must be stressed here that recall and precision ratios are not measures by which we can compare the performance of our system with that of some other system having different documents, different requests, and different user tolerances relating to recall, precision, response time, and amount of user effort. Such comparisons are meaningless.

Recall and precision ratios are essentially yardsticks. Within the context of an evaluation program, we use these yardsticks in the way that we use other yardsticks—namely, to measure things. The very least we can do is to measure the performance of the system, for the test requests, in relation to the ideal of 100% recall and 100% precision. More importantly, since the members of our test user group will have indicated varying requirements for recall and precision, we can measure how far the system has been able to meet these needs. Further, we can measure differences in performance for requests in various broad subject fields or for requests from various types of user groups. We can measure variations in performance when operating the system in alternative ways—for example, by experimenting with various levels and stages of interaction between the user and the system. We can also use these yardsticks to measure the effect of making changes to our system, such as the addition or omission of role indicators, the use of term weighting, and the interposition of a human intermediary to screen search output before delivery to end user.

When we consider that these ratios are merely tools by which we measure variations in performance within our own system, and within the confines of a controlled experiment, it is evident that any method that will give us reasonably accurate estimates of recall and precision is adequate, *so long as we hold the method constant throughout the evaluation program*. Even if the method results in slightly inflated, or slightly deflated, estimates of recall or precision, if it is held constant it will still result in performance figures that will be valid tools to use in the comparison of system alterations.

It must also be stressed that a test program is conducted merely to obtain data for analysis purposes. There are essentially two types of data that are produced in such a program: performance figures and "happenings."

We have already discussed performance figures and their uses. The "happenings" that we are primarily interested in, for analysis purposes, are instances of recall failures—*i.e.*, failures to retrieve known relevant documents—and precision failures—*i.e.*, failures to filter out nonrelevant documents. We want to know why these things happened. What were the factors contributing to these failures? Determination of these factors adversely affecting the performance of the system requires an examination of the documents involved, the indexing records for these documents, the request, the search formulation for this request, any clerical records available (coding sheets or Flexowriter hard copy), and the requester's relevance assessment sheets.

By an analysis of these records we can determine what factor, or factors, contributed to the recall and precision failures. Some of the major contributing factors are as follows:

Indexing.

(a) Careless mistakes of omission or use of an inappropriate term.
(b) Indexing insufficiently exhaustively, causing wanted documents to be missed; or too exhaustively, causing many unwanted documents to be retrieved.
(c) Failure to appreciate the importance of a particular concept due to subject weakness or lack of knowledge of the requirements of the user group.

Index Language.

(a) Lack of specificity, causing loss in filtering capacity.
(b) Overspecificity, causing loss of wanted documents. A common form of overspecificity is the use of roles, which are difficult to match in indexing and searching.
(c) More than one possible way to express the same idea: indexer and searcher chose different terms.
(d) False coordinations.
(e) Lack of sufficient hierarchical structure.
(f) Lack of sufficient linkages in the thesaurus.

Searching.

(a) Omission of important concepts demanded in the request.
(b) Use of terms more specific or less specific than appropriate.
(c) Lack of persistence in trying alternative searching strategies.

Clerical Failures.

(a) Keypunching errors.
(b) Coding errors.
(c) Journal issues missed in indexing

Computer Processing.

(a) Programs to check clerical failures break down.

Lack of Sufficient User-System Interaction.

On the basis of our analysis we can discover weak spots in our system, thus allowing corrective action to be taken. In the end, however, all matters reduce to a question of economics. Within the context of our evaluation, we may find that there are various alternative paths we can follow in order to produce the same end result to the system user. For example, there are at least three possible methods that we can employ in order to produce a final high-precision search product:

(1) Use of a sophisticated index language with syntactical-type controls in indexing—*e.g.*, links and roles.

(2) Use of a member of the information staff to screen output and weed out obviously nonrelevant material.

(3) Demanding more effort on the part of the user by means of increased interaction (at request, search formulation, or exploratory search stages) with the object of reducing the distance between his stated request and his real information need.

By testing we can determine which technique yields the best results in our own particular situation—in other words, which technique will produce a high precision search with minimum sacrifice of the other user requirements of recall, response time, and conservation of effort. Then it becomes a matter of pure economics: In relation to their relative performances, which method is the most economical and most feasible to implement in our own particular situation?

From a properly designed evaluation program, we can also expect to obtain payoff factors for certain aspects of our system. For example, we can derive cost/effectiveness ratios for various types of materials input to the system such as various report series or language groups, based on the number of times these materials are retrieved and judged of value related to the costs of inputting them.

Similarly, we can establish optimum levels of exhaustivity of indexing and optimum indexing times (levels above which the improvement in performance becomes insignificant when compared with the increased costs) for our own particular collection of documents in relation to the types of requests being made to it.

SUMMARY

We evaluate a system, large or small:

(1) To measure the degree to which the system is meeting user requirements.

(2) To locate sources of system failure, thereby allowing corrective action to be taken.

(3) To compare alternative methods of operating the system.

(4) To develop payoff factors for various aspects of the system.

We evaluate to allow improvement of the system either in its operating efficiency (which relates to degree of user satisfaction), or in its economic efficiency (which relates to optimum means of user satisfaction). System improvement in one or other, or both, of these aspects is the end result we can reasonably expect to achieve from a well-designed test program. System improvement is the sole justification for undertaking any evaluation program.

LITERATURE CITED

(1) Fairthorne, R. A., "Basic Parameters of Retrieval Tests," *Proc. Am. Documentation Institute Annual Meeting*, Philadelphia, Pa., October 5-8, 1964, p. 343.
(2) Aitchison, Jean, Cleverdon, Cyril, "A Report on a Test of the Index of Metallurgical Literature of Western Reserve University," The College of Aeronautics, Cranfield, England, October 1963.
(3) Johanningsmeier, W., Lancaster, F. W., *Project SHARP Information Storage and Retrieval System: Evaluation of Indexing Procedures and Retrieval Effectiveness*, Government Printing Office, Washington, D. C., 1964.