Chapter 5

Retrieval Evaluation

0 PREVIEW

Various automatic retrieval techniques were introduced in the last two chapters that have the potential to alter drastically the current operational retrieval environment. These techniques are not likely to find widespread favor unless their usefulness can be convincingly demonstrated. It is therefore important to understand the problems and techniques involved in the evaluation of retrieval systems and procedures.

This chapter is concerned with the evaluation of retrieval efficiency and effectiveness. Various viewpoints can be taken in evaluating a large system. The text stresses the user viewpoint and examines in detail the various system components which enter into the evaluation task. On the one hand, the system should be able to retrieve a large part of the relevant information contained in the files, while rejecting a large part of the extraneous information; on the other hand, the user effort, time, and cost needed for retrieval should be minimized. The former are characteristics of retrieval effectiveness, often measured by specific values of the recall and precision of the search output. The latter are components entering into an evaluation of search efficiency.

The generation and computation of the recall and precision measures are covered in detail in this chapter together with the computation of various alternative measures of retrieval effectiveness. The best-understood measure of system performance is the cost of using the system. Unfortunately, cost measures are difficult to generate for information retrieval systems. In the end, user satisfaction will depend on a multiplicity of considerations, including the ease with which the system retrieves wanted information, the cost of the system and amount of user effort required in effecting a search, and various human factors such as console design and physical location of the search equipment.

1 INTRODUCTION

To understand the retrieval evaluation problem, it is necessary to examine first the functions of a retrieval system and the various system components. Thereafter, the measures that actually reflect system performance can be introduced.

Information retrieval systems give a user population access to a stored collection of information items. These systems try to locate and retrieve the items as rapidly and economically as possible. The value of the information retrieval system depends on the ability to identify useful information accurately and quickly, the ease of rejecting extraneous items, and the versatility of the methods. Few customers will want a system incapable of retrieving what they want and of rejecting what they do not want. Nor will they want a system that is difficult to handle, slow in furnishing responses, or expensive to use. For these reasons, the evaluation of retrieval systems is of great importance [1–4].

Two kinds of system tests must be distinguished: those concerned with systems *effectiveness*, and those concerned with the *efficiency* of the operations. The effectiveness of an information system is the ability to furnish information services that the users need. On the other hand, efficiency is a measure of the cost or the time necessary to perform a given set of tasks. Ultimately, the viability of a system depends on both the quality and the cost of the operations. A complete evaluation process is then concerned with both effectiveness and efficiency.

There are many reasons for evaluating retrieval systems. For example, one might wish to compare one already existing system with another alternative system. One might also want to determine how system performance changes when some particular system component changes; for example, one could determine the performance changes when the query type is altered or when the subject area is changed. Still another reason is the evaluation of new system components that are considered for inclusion in an existing system. In that case, the operations of the new system could be simulated before a real system is actually built.

The following components are needed in a system test: (1) a detailed description of the system and of its components, or alternatively a model of the system to be examined; (2) a set of hypotheses to be tested, or a particular prototype against which the model is to be measured; (3) a set of criteria reflecting the performance objectives of the system, and measures permitting a quantification of the performance criteria; and finally (4) methods for obtaining and evaluating the data. For example, one might wish to look at the Dialog system and to determine whether the search speed is fast enough to search the data base, the criterion being that maximum allowable response time is 45 seconds, and the evaluation method consisting in asking each user of that data base to measure the response time with a stop watch.

To measure and record system performance, it is desirable to use objective, quantitative criteria. Objective measurements are relatively easy to interpret and are usually free of bias introduced by the evaluator. Objective measurements are obtainable by recording direct observations using questionnaire and interview techniques. Alternatively, some mechanical way may be used to gather the required data. In either case, parameters (that is, constants whose value characterizes the usefulness or worth of a system) must be chosen that are significant for evaluation purposes and are also easily measured and correlated. Some parameters are easily specified—for example, the size of the collection and the system response time. Many other important parameters are interdependent in complex ways—for example, the ability of the system to retrieve useful materials depends on the representation of the documents, the search methods, and the user characteristics. Further, some parameters may not be defined everywhere in the performance range. For example, the ability to retrieve useful items cannot be used when dealing with a document collection that contains no useful documents in a given subject [5–9].

In some circumstances, exact values may be unavailable for certain parameters, or they may be too laborious to supply. For example, the total number of documents in a particular subject area may be unknown and may have to be guessed at using sampling techniques. Probabilistic models are often applicable because many parameters become stable when many observations are made [10]. For example, the average number of relevant items per query or the average number of relevant items retrieved by the system can be estimated either when many documents are matched against a single request or by treating a few documents in many different searches.

2 EVALUATION OF RETRIEVAL EFFECTIVENESS

A System Components

Before a detailed examination of the evaluation parameters can be made, it is necessary to consider briefly the components of an information system and the system environment to determine how system performance is affected by the system environment and operations. The following system components are of concern: acquisitions and input policies, physical form of input, organization of the search files, indexing language, indexing operation, representation of the information items, question analysis, search, and form of presentation of the output [11,12].

Parameters related to the *input policies* include the error rates and time delays experienced in introducing new items into the collection, the time lag between receipt of a given item and its appearance in the file; and the collection coverage, that is, the proportion of potentially relevant information items actually included in the file. The *physical input form*, including document format and document length—title, abstract, summary, or full text—immediately affects the indexing and search tasks, as well as the system economics; and the *organization of the search files* impacts the search process, the response time, the effort needed by system operators, and possibly also the system effectiveness.

The *indexing language* consists of the set of available terms and the rules used to assign these terms to documents and search requests. During the indexing process terms appropriate for the representation of document content are chosen from the indexing language and assigned to the information items in accordance with established indexing rules. Among the parameters that take on special significance in this connection are the *exhaustivity* and *specificity* of the indexing language. An *exhaustive* indexing language contains terms covering all subject areas mentioned in the collection; correspondingly, an exhaustive indexing product implies that all subject areas are properly reflected in the index terms assigned to the documents. A *specific* index language never covers distinct subjects by using a single term, the terms used being narrow and precise.

Retrieval system performance is often measured by using *recall* and *precision* values, where recall measures the ability of the system to retrieve useful documents, while precision conversely measures the ability to reject useless materials. A high level of indexing exhaustivity tends to ensure high recall by making it possible to retrieve most potentially relevant items; at the same time precision may suffer because some marginally relevant items are likely to be retrieved also when many different subject areas are covered by the index terms. When highly specific index terms are used, the precision is expected to be high, since most retrieved items may be expected to be relevant; the converse is true when very broad or general terms are used for indexing purposes because broad terms will not distinguish the marginal items from the truly relevant ones. Thus to obtain high recall an exhaustive indexing is useful in conjunction with an indexing language that provides a variety of approaches to cover the given subject area. To ensure high precision, a highly specific indexing language should be used, and the terms should carry additional content indications such as term weights and relation indications to other terms.

Assuming that the indexing is performed manually by trained persons, the variables affecting the *indexing operation* relate not only to the exhaustivity of the indexing and the specificity of the assigned terms, but also to interindexer consistency, the influence of indexer experience on performance, and the accuracy of the assigned terms.

The *question analysis* and search operations are difficult to characterize. The assignment of terms from the indexing language to information requests, the formulation of meaningful Boolean statements, and the comparison of analyzed requests with the stored information are all complicated tasks. In principle, the content analysis operations are the same for documents and search requests, in the sense that the notions of exhaustivity and specificity are equally as applicable to queries as to documents. Thus, exhaustive query indexing using highly specific terms should produce maximum search recall and precision. In practice, the query processing is often quite distinct from the document indexing because the user is necessarily directly involved in the former but not the latter. In many systems, the query analysis and search operations are therefore delegated to trained experts using appropriate input from the users. Document input, on the other hand, is invariably handled without user input.

The *search operations* are also hard to measure using objective parameters because the role of the user is not well defined in many query formulating environments. Users are rarely asked to state recall or precision requirements, or to evaluate the output products. Yet search strategies need to be devised that respond to the users' specific recall and precision requirements. Among the characteristics that should be included in a measurement of search performance are the type of file organizations used, the type of query-document comparison in use, the effect of the search strategy on system response time and on search performance, and the relevance standards of the system users.

The *form of presentation of the output* is the physical representation of documents found by the system in response to the user's query. The appearance of the output affects the amount of user effort needed to look at the search results and the eventual satisfaction derived from a search. The more complete the form of the output, the easier is the relevance assessment task for the user. On the other hand, as the output is expanded from simple document numbers to full document texts, the time needed to examine the search results also increases.

The foregoing discussion makes clear that the components and parameters of the retrieval system affect the system operations and hence the evaluation results. Each component can be examined separately, or one can compare one entire system with another. In this case the parameters associated with each system are accepted as constant elements in the evaluation. However, one must understand that each of the parameters has an effect on the system, and the importance of each parameter cannot be assessed without taking into account the purposes for which the system is used. This question is discussed further in the next few paragraphs.

B Evaluation Viewpoints and the Relevance Problem

Information systems may be examined either from the viewpoint of the users or from the viewpoint of system operators and managers. For present purposes, the system managers may be assumed to include all those who influence the policy or the finances of the system, or who are responsible for, or participate in, the actual system operations. Since it is reasonable to assume that an information system exists to meet the needs of its users, the effectiveness criteria of interest to the managers are not unlike those of the users. In particular, the system should meet the user requirements, and failures in the retrieval of relevant materials or in the rejection of nonrelevant items should be minimized. In addition, the managers and to some extent the users are also concerned with the costs and benefits of the system.

Among the many possible evaluation criteria of concern to the user population, six have been identified as critical [13,14]:

1 The *recall*, that is, the ability of the system to present all relevant items
2 The *precision*, that is, the ability to present only the relevant items
3 The *effort*, intellectual or physical, required from the users in formulating the queries, conducting the search, and screening the output
4 The *time* interval which elapses between receipt of a user query and the presentation of system responses
5 The form of *presentation* of the search output which influences the user's ability to utilize the retrieved materials
6 The collection *coverage*, that is, the extent to which all relevant items are included in the system

A list of the criteria of interest to the user population is shown together with the principal related parameters in Table 5-1.

Table 5-1 User Performance Criteria and Related Parameters

User criteria	Selected related parameters
Recall and precision	Indexing exhaustivity (the more exhaustive, the better the recall) Specificity of indexing language (the more specific, the better the precision) Provisions in indexing language for improving recall (synonym recognition, recognition of term relations, etc.) Provisions in indexing language for improving precision (use of term weights, use of term phrases) Ability of user population to formulate search requests Ability to devise adequate search strategies
Response time	Type of storage device and storage organization Query type Location of information center Rate of arrival of customer queries Collection size
User effort	Characteristics of device permitting access to system Location of accessing and storage devices Availability of help from system staff or aids available from system in nondelegated searches Amount of retrieved material Type of interaction with system Ease of formulation of search requests
Form of presentation	Type of accessing and display device Size of stored information file Type of output (title, abstract, or full text)
Collection coverage	Type of input device and type and size of storage device Ease of content analysis (coverage may be more extensive when content analysis is simple) Demand for service (the demand increases with greater coverage)

Of the six user criteria, all but two are relatively easy to measure. The user effort can be expressed in part as the time needed for query formulation, the interaction with the system, and the examination of system outputs. The response time is directly measurable and the form of presentation of the output is easy to state. The collection coverage may present some difficulties if the number of items of interest in a given subject area is unknown. However, consulting published indexes and reference volumes should make it relatively easy to estimate the total number of items that are in fact available from the data base.

This leaves the recall and precision measures. These present the greatest difficulties both conceptually and in practice. An immediate problem in determining the recall and precision is the interpretation of *relevance*. At least two definitions of relevance are possible. An objective view takes into account only a given query and a particular document by stating that

> relevance is the correspondence in context between an information requirement statement (a query) and an article (a document), that is, the extent to which the article covers the material that is appropriate to the requirement statement [15].

That view makes it possible to consider relevance as a logical property between a pair of textual items. It is measurable by the degree to which a document deals with the subject of the user's information need [16,17]. The objective definition of relevance does not take into account the particular state of knowledge of the user during the search operation. A document might be "relevant" even though the user might already have been acquainted with the item before the formulation of the search request, or might have become familiar with the document through earlier search efforts.

A more subjective view of relevance considers not only the contents of a document but also the state of knowledge of the user at the time of the search, and the other documents retrieved or available that the user already knows about. Thus, this notion of relevance depends on the utility of each item to the user. The *pertinent* set of items may then be defined as that subset of the stored items that is appropriate to the user's information need at the time of retrieval [18,19]. Thus a document may be relevant if it deals with the appropriate topic classes but it may not be pertinent if the user is already acquainted with its contents, or if other documents retrieved earlier already cover the appropriate topics. All pertinent items are relevant but not vice versa.

Retrieval effectiveness may be easier to measure by using the objective view of relevance, or topic relatedness, as a criterion for determining relevance than the subjective notion of pertinence. Even in that case, difficulties may arise in assessing the relevance of a document to a query. In borderline cases disagreements may exist among observers about where to place the limit between various grades of relevance, and how to assess the relevance [20,21]. This has led some observers to define relevance in probabilistic terms. In this case relevance is a function of the probability that similarities between the query and document vocabularies will lead a user to accept a given item in response to a particular query [22].

In practice, it is necessary to assume that relevance assessments of documents to queries are available from an external source to the retrieval system if an objective system evaluation is to be accomplished. Hence, a system is judged to be effective if satisfactory evaluation results are obtained using the external relevance criteria.

In the next few sections problems relating to the computation and presentation of the recall and precision measures are examined and alternative retrieval evaluation measures are introduced.

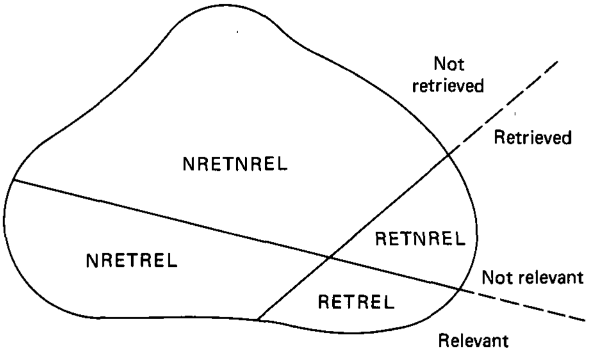
*C The Computation of Recall and Precision

Recall is defined as the proportion of relevant material retrieved, while precision is the proportion of retrieved material that is relevant. In an operational situation, where information needs may vary from user to user, some customers may require high recall, that is, the retrieval of most everything that is likely to be of interest, while others may prefer high precision, that is, the rejection of everything likely to be useless. Everything else being equal, a good system is one which exhibits both a high recall and a high precision.

If a cut is made through the document collection to distinguish retrieved items from nonretrieved ones on the one hand as shown in Fig. 5-1, and if procedures are available for separating relevant items from nonrelevant on the other, the standard recall R and standard precision P may be defined as

$$R = \frac{\text{number of items retrieved and relevant}}{\text{total relevant in collection}} \tag{1}$$

$$\text{and } P = \frac{\text{number of items retrieved and relevant}}{\text{total retrieved}} \tag{2}$$



RETREL : number relevant and retrieved
RETNREL : number not relevant and retrieved
NRETREL : number relevant and not retrieved
NRETNREL: number not relevant and not retrieved

Figure 5-1 Partition of collection.

If relevance judgments are available for each document in the collection with respect to each search request, and if retrieved and nonretrieved material can be unambiguously determined, then the computation of these measures is straightforward [23,24].

In conventional retrieval systems the search requests are presented as Boolean combinations of search terms. The retrieved document set consists of all documents exhibiting the exact combination of keywords specified in the query. That is, each query produces an unordered set of documents that are either relevant or nonrelevant. Hence for each query a single precision and a single recall figure can be obtained. Pairs of recall-precision figures can be compared for two searches i and j, and whenever $\text{RECALL}_i \leq \text{RECALL}_j$ and $\text{PRECISION}_i \leq \text{PRECISION}_j$ the results of search j are judged to be superior to those for search i. Unfortunately, problems arise when $\text{RECALL}_i < \text{RECALL}_j$ and $\text{PRECISION}_i > \text{PRECISION}_j$ or vice versa $\text{RECALL}_i > \text{RECALL}_j$ and $\text{PRECISION}_i < \text{PRECISION}_j$ [25]. In these cases, a judgment of superiority depends on the user's orientation. That is, the user must determine if the principal interest is in recall or in precision, and assess the importance of differences between the recall and precision values. In typical retrieval systems, the recall will increase as the number of retrieved documents increases; at the same time, the precision is likely to decrease. Hence users interested in high recall tend to submit broad queries that retrieve many documents, whereas high-precision users will submit narrow and specific queries.

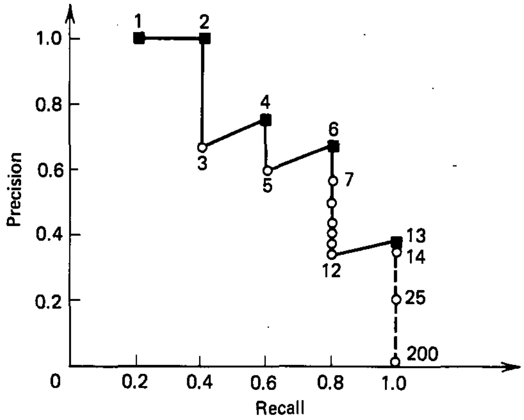
Some retrieval systems can produce varying amounts of output. A different recall-precision pair can then be obtained for each separate output amount. The finer the division in quantity of output, the greater the number of available recall-precision pairs. For example, the retrieval decision can be based on the number of matching terms between queries and documents. A partial ranking can then be defined for the retrieved document set by first retrieving all items that exhibit at least some arbitrary k matching terms with the query for some judiciously chosen number k. Next all items with $k - 1$ matching terms are retrieved, followed by those with $k - 2$ matching terms, and so on down to the items that have no terms in common with the query. In each case the greater the number of matching query terms the higher the rank of the document in the list of retrieved documents. In such a system several different pairs of recall-precision values can be computed depending on the number of matching terms between queries and documents.

In a number of retrieval systems, a ranking is obtained for the retrieved document set by computing a similarity coefficient for each document-query pair. This coefficient reflects the similarities between the corresponding index terms or content representation. The retrieved items are then listed in decreasing order of the query-document similarity coefficients [24]. A pair of recall-precision values can then be computed following the retrieval of each document in the ranked order.

The recall measurement requires knowledge of the total number of relevant documents in the collection with respect to each query. When the size of the document collection is relatively small, it is often possible to obtain relevance judgments for all documents with respect to each query. When the collection sizes are larger, such exhaustive relevance assessments are not normally available. To obtain dependable recall figures, it is then necessary to estimate the total number of relevant documents in the collection. This can be done by sampling techniques. Thus, relevance assessments are made for only a subset of the items in a collection [26]. Alternatively, a given query could be processed using a variety of different search and retrieval methods with the assumption that all relevant documents are probably going to be retrieved by the various searches. The results of the searches are then combined into a single

		Recall-precision after retrieval of n documents	
n	Document number (x = relevant)	Recall	Precision
1	588 x	0.2	1.0
2	589 x	0.4	1.0
3	576	0.4	0.67
4	590 x	0.6	0.75
5	986	0.6	0.60
6	592 x	0.8	0.67
7	984	0.8	0.57
8	988	0.8	0.50
9	578	0.8	0.44
10	985	0.8	0.40
11	103	0.8	0.36
12	591	0.8	0.33
13	772 x	1.0	0.38
14	990	1.0	0.36

(a)



(b)

Figure 5-2 Display of recall and precision results for a sample query. (Collection consists of 200 documents in aerodynamics.) (a) Output ranking of documents in decreasing query-document similarity order and computation of recall and precision values for a single query. (b) Graph of precision versus recall for sample query of Fig. 5-2a.

output list. The list of relevant documents is obtained following a relevance assessment of this single output list.

Consider, as an example of a recall-precision computation, a query which has a total of five relevant documents included in a collection of two hundred documents in aerodynamics. The ranks of the relevant items in decreasing query-document similarity order are shown in Fig. 5-2a as well as the recall and precision values based on these ranks [27]. The recall-precision values are obtained from equations (1) and (2). So if six documents are retrieved including four of the possible five relevant documents, then this produces a recall of $^4/_5$, or 0.8. The precision is computed as $^4/_6$ (4 relevant out of 6 retrieved), or 0.67.

Given a set of recall-precision value pairs, such as that in Fig. 5-2a, a recall-precision graph can be constructed by plotting the precision against the recall. The graph for the sample query of Fig. 5-2a is shown in Fig. 5-2b.

Recall-precision graphs, such as that of Fig. 5-2b, have been criticized because a number of parameters are obscured. For example, the size of the retrieved document set and the collection size are not available from the graph [28]. Furthermore, problems arise when producing a continuous graph from a discrete set of points. That is, the value of precision is known exactly for a recall of 0.2 in the example of Fig. 5-2a (1.0), but it is not exactly specified for a recall of 0.4, since the precision varies between 1.0 and 0.67 at that point. Similarly the recall value is specified exactly when the precision is 0.5, but not when it is 1.0. Another problem arises when a number of curves such as the one of Fig. 5-2b, each valid for a single query, must be processed to obtain average performance characteristics for many user queries.

Before defining a single composite recall-precision graph reflecting the average performance of a system for a large number of individual queries, it is convenient first to replace the sawtooth curves for the individual queries, by smoother versions that simplify the averaging process. One possibility consists in using graphs consisting of horizontal line segments such as those shown in Fig. 5-3 for the example of Fig. 5-2. The curve of Fig. 5-3 is obtained by starting

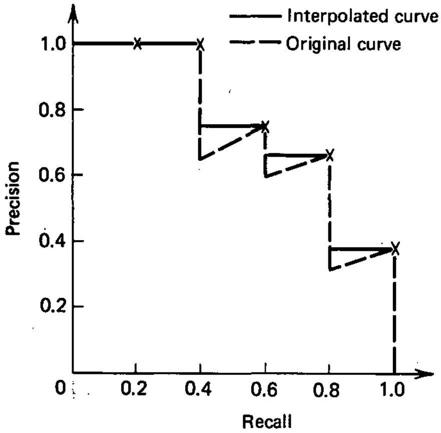


Figure 5-3 Interpolated recall-precision curve for sample query of Fig. 5-2. (Ranks of relevant items are 1, 2, 4, 6, 13.)

at the highest recall value and drawing a horizontal line leftward from each peak point of precision, up to a point where a higher precision point is encountered. The curve of Fig. 5-3 now exhibits a unique precision value for each recall point, and it extends along the scale from a recall of 0 to a recall of 1. For example, at a recall of 0.4 the precision is 1.0 in the graph of Fig. 5-3; however, for any slightly larger value of the recall—say 0.401—the precision has dropped to 0.75. A similar drop in precision from 0.75 to 0.67 occurs as the recall increases from 0.6 to 0.601. The interpolated curve represents the best performance a user can achieve [27].

Given a set of different performance (recall-precision) curves similar to that of Fig. 5-3, corresponding to different user queries, average performance values can be obtained in several different ways. In particular, if RETREL_i is defined as the number of items retrieved and relevant, RETNREL_i is the number retrieved but not relevant, and NRETREL_i is the number relevant but not retrieved for query i, then following the definitions in (1) and (2), the RECALL_i for query i, and the PRECISION_i are defined as

$$\text{RECALL}_i = \frac{\text{RETREL}_i}{\text{RETREL}_i + \text{NRETREL}_i} \tag{3}$$

$$\text{and PRECISION}_i = \frac{\text{RETREL}_i}{\text{RETREL}_i + \text{RETNREL}_i} \tag{4}$$

A *user-oriented recall-level average*, reflecting the performance an average user can expect to obtain from the system, may then be defined by taking the arithmetic mean, over NUM sample queries, of expressions (3) and (4):

$$\text{RECALL}_{\text{RL}} = \frac{1}{\text{NUM}} \sum_{i=1}^{\text{NUM}} \frac{\text{RETREL}_i}{\text{RETREL}_i + \text{NRETREL}_i} \tag{5}$$

$$\text{PRECISION}_{\text{RL}} = \frac{1}{\text{NUM}} \sum_{i=1}^{\text{NUM}} \frac{\text{RETREL}_i}{\text{RETREL}_i + \text{RETNREL}_i} \tag{6}$$

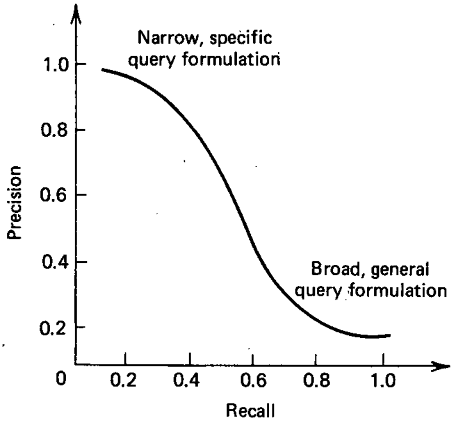


Figure 5-4 Typical average recall-precision graph.

Since the recall and precision values RECALL_i and PRECISION_i for the individual user queries are unambiguously defined as shown in Fig. 5-3, the averages of equations (5) and (6) are also uniquely determined. This makes it possible to compute average precision values at fixed recall intervals, say for recall equal to 0, 0.1, 0.2, . . . , 1.0. In particular, for each query the precision values are computed for the specified 11 values of the recall from 0 to 1.0 in steps 0.1, and equation (6) is used to obtain average precision values over all queries at each of the 11 recall values. The average curve which results has a shape similar to that shown in Fig. 5-4, where the left-hand end corresponds to narrow, specific query formulations where few documents are retrieved and the precision may be expected to be high, while the recall is fairly low. The right-hand end of the curve represents broad, rather general query formulations and hence a large number of retrieved documents.

An alternative *systems-oriented document-level average* is obtained by using the total number of relevant items retrieved by the system over the NUM queries, as well as the total number of nonrelevant items that are rejected. That is, from the NUM original queries, a single hypothetical combined query is built, whose relevant items are defined as the sum of the relevant of all component queries. The document level averages are then defined as

$$\text{RECALL}_{\text{DL}} = \frac{\sum_{i=1}^{\text{NUM}} \text{RETREL}_i}{\sum_{i=1}^{\text{NUM}} (\text{RETREL}_i + \text{NRETREL}_i)} \tag{7}$$

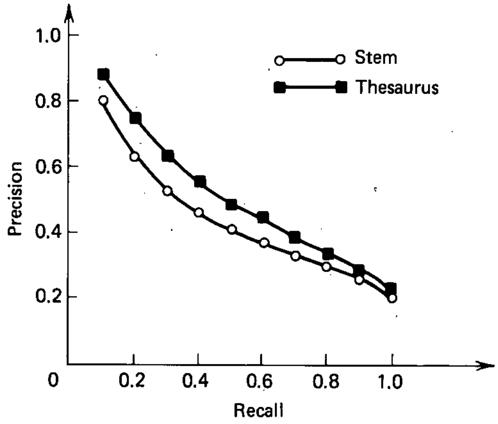
$$\text{and PRECISION}_{\text{DL}} = \frac{\sum_{i=1}^{\text{NUM}} \text{RETREL}_i}{\sum_{i=1}^{\text{NUM}} (\text{RETREL}_i + \text{RETNREL}_i)} \tag{8}$$

The averages of equations (5) and (6) give equal importance to each query, while in formulas (7) and (8) the averages depend more on queries with many relevant documents than on those with few relevant items. Consider by analogy a computation of average class size in a university. If there are 10 classes, including 5 with 1 student each and 5 with 99 students each, the *class-level* average size is 50, reflecting the fact that 10 professors teach a total of 500 students, or 50 on average. The *student-level* average size, on the other hand, is 98.02, reflecting the fact that almost all students are in classes with 98 other students. In information retrieval the choice of averaging method depends on whether it is more important to display the average user's result [equations (5) and (6)] or to reflect what happens to the average relevant document [equations (7) and (8)]. If query performance does not depend on the number of relevant documents, the two averages give similar results.

Recall-precision curves may be used to evaluate the performance of information retrieval systems—typically, by computing recall and precision values for two or more systems, or for the same system operating under different con-

Recall	Average precision for 35 queries		Improvement, %
	Word stem	Thesaurus	
0.1	0.7963	0.8788	10.4
0.2	0.6350	0.7567	19.2
0.3	0.5283	0.6464	22.4
0.4	0.4603	0.5577	21.2
0.5	0.4051	0.4912	21.3
0.6	0.3699	0.4470	20.8
0.7	0.3383	0.3893	15.1
0.8	0.2996	0.3287	9.7
0.9	0.2568	0.2726	6.2
1.0	0.2018	0.2093	3.7

(a)



(b)

Figure 5-5 Average recall-precision results for two indexing methods (82 documents, 35 queries). (a) Recall-precision average. (b) Recall-precision graph.

ditions. In these circumstances, the curves produced for systems A and B can be superimposed on the same graph to determine which system is superior, and by how much. In general, the curve closest to the upper right-hand corner of the graph (where recall and precision are maximized) indicates the best performance. A typical example is shown in Fig. 5-5, where the performance of two different indexing systems is shown for a collection of documents in library science averaged over 35 user queries. The "stem" run refers to an indexing process where word stems extracted from the document abstracts are used as index terms to represent document content. In the "thesaurus" run word stems are replaced by "concepts" extracted from a thesaurus representing classes of terms related or synonymous to the original stems. It may be seen from the output of Fig. 5-5 that the average precision of the thesaurus run is between 4 and 22 percent better at the fixed recall points than the word stem run.

Since it is difficult to judge the significance of the differences between two performance curves by citing percentage improvements as in Fig. 5-5, it is helpful to furnish statistical evidence indicating whether a given difference between two averages is in fact significant. Most standard *statistical significance tests* based on paired comparisons will produce statistical evidence giving the probability that differences between the two sets of sample values as great as, or greater than, those observed would occur by chance. When the computed probability is small enough—for example, less than or equal to 0.05—one concludes that the two sets of sample values are significantly different. If, on the other hand, the computed probability is greater than 0.05, the presumption is that the observed differences could have been obtained by chance—that the original pairs of values might in fact have been derived from the same distribution.

The pairs of measurements being compared may typically represent the precision values at some fixed recall level—say, at a recall of 0.1, or at a recall of 0.5—for a set of queries processed by using methods A and B, respectively. The two middle columns of Fig. 5-5a represent an example of this case. Alternatively, the pairs of measurements may represent combined values for several points representing the complete recall-precision curves. Such combined measurements are obtained from the single-valued evaluation measures introduced later in this chapter.

The following assumptions are made for three of the best-known significance testing procedures [29]. The measurements must be obtained independently of each other;

1 For the *t-test* it is assumed in addition that the differences between the two sets of sample values to be compared are normally distributed.

2 The *sign test* makes no normality assumptions, and uses only the sign (not the magnitude) of the differences in sample values; thus the computed probability values depend on whether the differences in sample values are mostly positive or negative.

3 The *Wilcoxon signed rank test* postulates that as differences between pairs increase, significance also increases, but only as these numbers affect the ranking; thus, differences of $-1, 2, -3, 4$, and 20 are equivalent to differences of $-1, 2, -3, 4$, and 5, since only the rank of the ordered differences is important.

Since many sets of recall or precision differences are probably not normally distributed, the less stringent assumptions inherent in the use of the sign test or the Wilcoxon signed rank test may be preferable over those of the better known t-test. A typical set of output data from a sign test process is shown for two search methods A and B in Table 5-2. The table is based on 11 statistics (differences in precision at each of 11 recall values from 0 to 1 in steps of 0.1). For each statistic this table shows the number of queries favoring methods A and B, respectively, and the one-sided probabilities for the test (ignoring ties). The one-sided probabilities represent the probabilities that the sample values could have originated by chance. On the bottom line of Table 5-2 the 11 one-sided tests are combined into a single overall measure. In this case the probabilities measure the chance that method B is not significantly different (better)

Table 5-2 Sign Test for Typical Search Methods A and B (Average for 42 Queries; Testing for Collection B Better than for Collection A)

Precision average at recall of	Favoring method A	Favoring method B	Favoring method Tied	One-sided probabilities
0	1	7	34	0.0385
0.1	1	7	34	0.0385
0.2	1	8	33	0.0226
0.3	1	13	28	0.0018
0.4	1	15	26	0.0006
0.5	2	17	23	0.0007
0.6	3	20	19	0.0004
0.7	8	16	18	0.0767
0.8	9	17	16	0.0851
0.9	8	18	16	0.0387
1.0	8	18	16	0.0387
Combined	43	156	263	0.0000

Adapted from reference 29.

than A. This is seen to be zero to four decimal places; that is, method B is statistically better than method A.

A majority of the studies undertaken to evaluate the effectiveness of information retrieval systems have used recall and precision measurements to show system performance. Misgivings have been voiced about some of the characteristics of these measures. As a result a variety of alternative methodologies have been proposed. A few developments in this area are examined in the next section.

3 MEASURES OF RETRIEVAL EFFECTIVENESS

A Measurement Problems

The recall and precision measures introduced earlier are advantageous to the user because they reflect the relative success of the system in meeting various kinds of user needs. Furthermore a particular measurement can be directly interpreted in terms of user experiences. Thus, a precision performance of 0.2 at recall equal to 0.5 implies that the user has obtained one-half of the relevant items in the collection, and that four nonrelevant items have had to be examined for every relevant item that was obtained.

Some qualifying remarks are nevertheless in order in connection with the standard recall and precision measurements. First, it is clear from the basic definitions that recall and precision measurements are normally tied to a given collection of documents and to a given query set. Within such a fixed environment, it is possible to vary the indexing policy or the indexing language or the search methodology and to determine subsequently how these changes may affect the performance of the system in terms of recall and precision. On the other hand, recall and precision must be used with caution in comparing the performance of two entirely different systems based on different document collections, different query sets, and different user populations [30,31].

Consider in particular the changes to be expected in the value of the recall and precision measures when the collection size increases or when the average number of relevant items per query diminishes from one collection to another. In both cases recall and precision can be expected to deteriorate because the number of relevant and retrieved items is not likely to increase in proportion to the size of the collection. Care must then be taken to equate collection and query relevance properties before applying recall and precision measures to the evaluation of different collections.

Another problem arises in connection with the relevance assessments of documents with respect to user queries. Such assessments are needed if the relevant items are to be distinguished from the nonrelevant ones. Some observers maintain that recall and precision measurements apply only to the user environment within which the relevance judgments were first obtained. This is because of the inherent subjectivity of relevance assessments [32]. Fortunately, there exists a good deal of experimental evidence to show that for many of the documents that appear to be most similar to a particular query, and hence are normally retrievable by the search process, very close agreement may be expected from different assessors as to the relevance in each case. This accounts for the fact that while the relevance assessments obtainable from differing evaluators are different for randomly chosen documents, the effect on the resulting recall and precision measurements is relatively small [33]. Furthermore, it is possible to replace the individual opinions about the usefulness of a given document with respect to a given query by global judgments representing a consensus of ideas by several independent judges [34,35].

A third question of interest in using recall and precision measurements is the effect of the query type on the evaluation outcome. In this respect one can distinguish the short *subject-heading* queries in which the topic is expressed as a single short descriptive word string (e.g., "thermal control," "turbulence studies") from *title-length* queries, where a single sentence or title adequately describes the subject area, and from *full-text* queries where a complete paragraph is used to formulate a search request [34]. Different query types may be produced notably in systems where the final query formulations are delegated by the users to trained search intermediaries. Although the short, subject-heading queries will often deal with general topics, whereas the larger full-text queries are sometimes more specific, it is not always true that query length is directly correlated with query specificity. In any case, a system should be tested using a realistic query mix reflecting the query types actually submitted in operational situations.

A last consideration relating to recall and precision computations is the assignment of relevance grades to the documents of a given collection, and the choice of a document rank for output purposes. Under normal circumstances, two relevance assessments are customary: either a document is relevant or it is nonrelevant. In these circumstances, the computation of recall and precision is unambiguous using expressions (1) and (2). If a system uses grades of relevance and the retrieved documents are ranked, several documents may be equally similar to a given query and placed in consecutive location on the output list. However, since the order affects the precision-recall evaluation, techniques are required to compensate for the arbitrary ordering of the equally similar items. One technique is to assign to all these documents a relevance grade equal to the average grade of this set of documents.

In practice, it appears to be a great deal more difficult for the relevance assessors to use many relevance grades than to simply decide between the relevant or partly relevant documents on the one hand and the nonrelevant documents on the other. Furthermore, errors and uncertainties crop up with multiple-category relevance assessments where users are forced to make narrow distinctions between documents that are absolutely relevant, possibly relevant, marginally relevant, or nonrelevant as the case may be. These in fact may outweigh the greater accuracy sought by using the many relevance grades. Nevertheless, various evaluation procedures and parameters have been proposed for use with variable relevance weights and for systems allowing ties in the ranks of the retrieved items. These measures are introduced with additional evaluation criteria in the next section.

*B Recall, Precision, and Fallout

It was pointed out earlier that recall and precision measurements are directly interpretable by users in terms of search satisfaction. On the other hand, they are sometimes difficult to compute. For example, recall may not be defined because no relevant documents exist in a collection with respect to some query [28,36]. Similarly, precision is undefined when no items are retrieved (the respective measures are computed as 0/0 in each case, which is undefined).

Another deficiency in the use of precision and recall is a lack of parallelism in the properties of the two measures. Assuming that retrieval effectiveness increases with the number of relevant items obtained in answer to a query, and decreases with the number of nonrelevant items retrieved, a measure appears to be needed which reflects the performance for the nonrelevant documents in the same way as recall measures the performance of the relevant. This measure, known as*fallout*, is formally defined using the terminology of Table 5-3 as

$$\begin{aligned}\text{FALLOUT} &= \frac{\text{RETNREL}}{\text{RETNREL} + \text{NRETNREL}}\\ &= \frac{\text{number of nonrelevant items retrieved}}{\text{total number of nonrelevant in collection}} \qquad (9)\end{aligned}$$

If the recall is expressed in probabilistic terms as the probability of a document being retrieved given that it is relevant, the fallout is the probability of an item being retrieved given that it is nonrelevant. An effective retrieval system will therefore exhibit maximum recall and minimum fallout.

In a normal retrieval environment, recall, precision, and fallout are not independent of the *generality factor*, defined as the average number of relevant

Table 5-3 Contingency Table

	Relevant	Nonrelevant	
Retrieved	RETREL	RETNREL	RETREL + RETNREL
Not retrieved	NRETREL	NRETNREL	NRETREL + NRETNREL
	RETREL + NRETREL	RETNREL + NRETNREL	RETREL + RETNREL + NRETREL + NRETNREL

items per query included in the collection. Referring to Tables 5-3 and 5-4, one sees that any three of the measures R, P, F, and G automatically determine the fourth. As an example, precision may be determined in terms of recall, fallout, and generality as

$$P = \frac{R \cdot G}{(R \cdot G) + F(1 - G)} \tag{10}$$

However, because the total number of nonrelevant items (RETNREL + NRETNREL) is much larger in practice than the number of relevant (RETREL + NRETREL), any changes in the generality of a collection are likely to affect the fallout less than the precision. In particular, as the generality decreases either because the number of relevant items decreases or because the total collection size increases, the number of relevant retrieved (RETREL) is likely to decrease, but the total number of items retrieved (RETREL + RETNREL) as well as the number of nonrelevant items (RETNREL + NRETNREL) may remain fairly constant. Hence precision will be subject to larger variations than fallout [31]. Furthermore, fallout is unequivocally defined to be zero when no items are retrieved (RETREL + RETNREL = 0), because the number of nonrelevant items in the collection (RETNREL + NRETNREL) may safely be assumed to be nonzero.

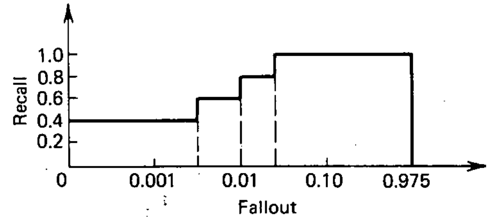
These arguments have been used to suggest the replacement of the recall and precision measures by recall-fallout computations. A typical recall-fallout display is shown in Fig. 5-6 for the query previously used as an example in Figs.

Table 5-4 Typical Retrieval Evaluation Measures

Symbol	Evaluation measure	Formula	Explanation
R	Recall	$\frac{\text{RETREL}}{\text{RETREL + NRETREL}}$	Proportion of relevant actually retrieved
P	Precision	$\frac{\text{RETREL}}{\text{RETREL + RETNREL}}$	Proportion of retrieved actually relevant
F	Fallout	$\frac{\text{RETNREL}}{\text{RETNREL + NRETNREL}}$	Proportion of nonrelevant actually retrieved
G	Generality	$\frac{\text{RETREL + NRETREL}}{\text{RETREL + RETNREL + NRETREL + NRETNREL}}$	Proportion of relevant per query

n	Relevant	Recall	Fallout
1	x	0.2	0
2	x	0.4	0
3		0.4	0.005
4	x	0.6	0.005
5		0.6	0.010
6	x	0.8	0.010
7		0.8	0.015
8		0.8	0.020
9		0.8	0.025
10		0.8	0.030
11		0.8	0.035
12		0.8	0.040
13	x	1.0	0.040
14		1.0	0.045
20		1.0	0.075
50		1.0	0.225
100		1.0	0.475
200		1.0	0.975

(a)



(b)

Figure 5-6 Display of recall and fallout results for sample query of Fig. 5-2. (Collection consists of 200 documents in aerodynamics.) (a) Recall-fallout after retrieval of n documents. (b) Recall-fallout plot.

5-2 and 5-3. The graph of Fig. 5-6 indicates that the fallout is not as easily interpreted by the user as precision. In fact, the two types of effectiveness pairs (recall-precision and recall-fallout, respectively) may well respond to different needs in actual retrieval situations. Since the recall provides an indication of the proportion of relevant actually obtained as a result of a search, while precision is a measure of the efficiency with which these relevant are retrieved, a recall-precision output is *user-oriented*, because the user is normally interested in optimizing the retrieval of relevant items. On the other hand, fallout is a measure of the efficiency of rejecting the nonrelevant in the collection (which in many cases is approximately equivalent to the collection size). For this reason, a recall-fallout display may be considered to be *systems-oriented*, since it indicates how well the nonrelevant are rejected as a function of collection size [37].

The foregoing measures are all based on objective relevance judgments that are independent of the user's prior knowledge of the subject area. Additional measures which depend on the subjective relevance are [27]:

1 The *novelty ratio,* that is, the proportion of items retrieved and judged relevant by users of which they had not been aware prior to receiving the search output

2 The *coverage ratio*, that is, the proportion of relevant items retrieved out of the total relevant known to users prior to the search
3 The *sought recall*, defined as the total relevant examined by users following a search, divided by the total relevant which users would have liked to examine

Many other evaluation measures based on the contingency table display of Table 5-3 have been proposed over the years [19,28,38,39]. Some of these measures use the full information incorporated in the contingency table, as opposed to recall, precision, and fallout that are based on a single column or a single row of the table only. However, most of these measures are not easily interpretable, and it is not likely in these circumstances that they will quickly supplant recall and precision.

**C Single-Valued Measures

Some observers completely reject the contingency table as a basis for the construction of parameters capable of reflecting retrieval effectiveness. Instead a number of desirable properties are postulated for an ideal effectiveness measure, including in particular the following [40]:

1 The measure should be able to reflect retrieval effectiveness alone, separately of the criteria such as cost.
2 The measure should be independent of any particular retrieval cutoff, that is, of the number of documents retrieved in a particular search.
3 The measure should be expressible as a single number (instead of two values such as recall and precision) which can be put on a scale to give absolute and relative values.

The best known of these single-valued measures is the E measure introduced by Swets [40–43].

To construct the E measure, two distinguishable populations of objects POP_1 and POP_2 are associated with the relevant and the nonrelevant documents with respect to some query. If a parameter r is used to represent some measurable characteristic such as the query-document similarity for each document, the "probability density functions" $\mathrm{FUNC}_1(\mathrm{r})$ and $\mathrm{FUNC}_2(\mathrm{r})$ with means MEAN_1 and MEAN_2 and variances VAR_1 and VAR_2 will then indicate how the parameter r behaves for the two populations. That is, $\mathrm{FUNC}_1(\mathrm{r})$ represents the probability that a document in POP_1 has value r, and $\mathrm{FUNC}_2(\mathrm{r})$ applies similarly to POP_2. A typical graph of these probability density functions is shown in Fig. 5-7.

By choosing a particular value of r—for example, the point $\mathrm{r} = \mathrm{C}$—it becomes possible to compute the fraction of each population for which r has a value greater than or equal to C. This is the portion of the area under the probability density curve lying to the right of line $\mathrm{r} = \mathrm{C}$. C is in fact a cutoff value such that any item for which $\mathrm{FUNC}(\mathrm{r}) \geq \mathrm{C}$ is retrieved. Since $\mathrm{FUNC}_1(\mathrm{r})$ and $\mathrm{FUNC}_2(\mathrm{r})$ are associated with the relevant and nonrelevant document popula-

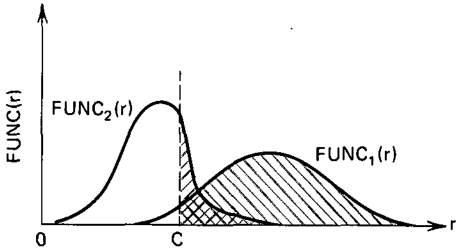


Figure 5-7 Probability density functions for populations POP$_1$ and POP$_2$ comprising the relevant and nonrelevant documents, respectively. (*Adapted from reference 41.*)

tions, respectively, the areas under the density curves to the right of r = C represent, respectively, the proportion of relevant and of nonrelevant documents for which FUNC(r) ≥ C. The first measure is equal to the recall, while the second is equal to the fallout. By plotting the percentages of the populations POP$_1$ and POP$_2$ to the right of the cut C against each other, while varying C, one obtains an *operating characteristic* (OC) curve similar to that shown in Fig. 5-8. If populations POP$_1$ and POP$_2$ are identical with respect to the characteristic r, the operating characteristic is a line running diagonally across the graph. The more different the two populations are from one another the more closely the OC curve will approach the upper left-hand corner of the figure (the 0-100 point of the graph).

Swets plotted the operating characteristics for a large number of retrieval systems using normal probability scales for recall and for fallout. In these circumstances, straight lines are obtained if recall and fallout both show normal distributions with respect to r. Within the limits of experimental error, the lines were all found to be straight, leading to the conclusion that the probability density functions of recall and fallout with respect to parameter r are normal. It follows that recall-fallout performance can be represented by specifying the po-

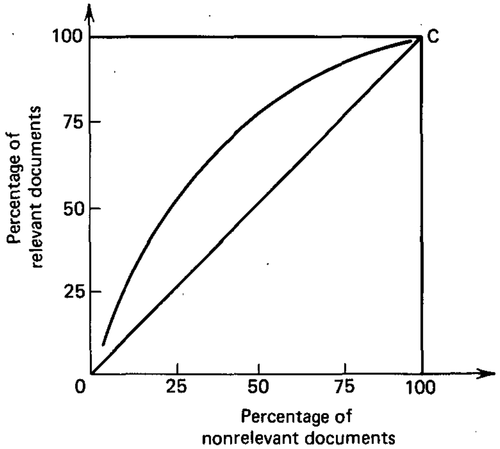


Figure 5-8 Operating characteristic curve. (*Adapted from reference 41.*)

sition of the corresponding straight line. Two typical operating characteristic lines are shown in Fig. 5-9, labeled A and B respectively. Swets' E measure is defined specifically as

$$E = \sqrt{2} \cdot \text{DIST} \tag{11}$$

where DIST is the distance from 0 (see Fig. 5-9) to the operating characteristic curve along the line from point 0 to point R. If the angle of the operating characteristic curve is the same as the diagonal, as it is for line A of Fig. 5-9, that is, the slope of the curve is equal to 1, then the value of E does represent the performance effectiveness. If the angle of the operating characteristic curve is not the same as the diagonal as in line B, then it is necessary to present the slope of the operating characteristic curve as well as the E value.

The main advantage of the (E, SLOPE) measures is that they are derived using a well-known and accepted statistical theory. The disadvantage is that unlike recall and precision, the E and SLOPE measures cannot be translated into a performance characterization as readily by the user population. In practice it is found that values of SLOPE range from 0.5 to 2.0, so that the two values (E, SLOPE) are necessary to express the performance, just as is the case for recall and precision. It should be noted also that the determination of the E measure is based on information equivalent to that contained in a full recall-precision graph. That is, the E value cannot be obtained by using a single retrieval threshold that distinguishes the retrieved from the nonretrieved items of the kind normally available in conventional retrieval situation; a single pair of recall-precision values is, however, computable in that situation.

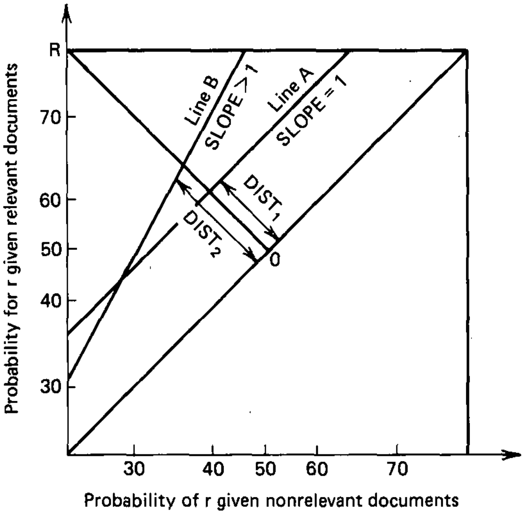

R
70
60
50
40
30
Probability for r given relevant documents
Line B
SLOPE > 1
Line A
SLOPE = 1
DIST1
DIST2
0
30
40
50
60
70
Probability of r given nonrelevant documents

Figure 5-9 Operating characteristics on normal probability scales. (*Adapted from reference 40.*)

Various additional global measures based on established theories—especially probability theory and information theory—have been described in the literature but have not received consideration in practice [44–49]. These measures generally combine aspects of precision and recall in a single expression. One function, based on considerations of measurement theory, uses a special parameter (α) which makes it possible to attach degrees of importance to the recall and precision components [25]:

$$E = 1 - \frac{1}{\alpha(1/\mathrm{PRECISION}) + (1 - \alpha)(1/\mathrm{RECALL})} \tag{12}$$

Large values of the recall and precision measure correspond to small values of the evaluation measure E. For example, assume an α value of 0.50 and recall and precision values equal to 0.50. These parameters produce an E value of 0.50. When the recall remains at 0.50 but precision drops to 0.25, the E value increases to 0.67. On the other hand, if recall remains at 0.50 and precision increases to 0.90, the E value is 0.36.

One composite evaluation measure is independent of the retrieval threshold used to distinguish the retrieved from the nonretrieved items and is applicable to systems that rank the retrieved documents. It is based on considerations similar to those used by Swets, in that the area is computed under a particular form of the recall and precision graph [50,51]. Consider a graph in which the recall of a system is represented along the ordinate and the rank orders of the retrieved documents are plotted along the abscissa. The graph starts at zero and maintains a zero value until a rank is reached corresponding to a retrieved relevant document, at which point the recall jumps to 1/REL (for REL relevant documents included in the system). The recall then stays at 1/REL until the next relevant document is reached, at which point the recall increases to 2/REL, and so on, until the last relevant document rank is reached when the recall reaches its final value of REL/REL or 1.

If this recall step function is plotted on the same graph with a similar function for an ideal system for which the REL relevant documents are ranked 1,2, . . . ,REL, the area between the two step functions can be used as a measure of the recall performance of the system. A typical case is shown in Fig. 5-10 for the query used as an example in Figs. 5-2, 5-3, and 5-6. The ranks of the relevant documents in decreasing similarity order are assumed to be 1, 2, 3, 4, 6, and 13. The difference in area between the ideal retrieval situation and the actual recall curves designated by the tinted area in the example in Fig. 5-10 is given by

$$\frac{\sum_{i=1}^{\mathrm{REL}} \mathrm{RANK}_i - \sum_{i=1}^{\mathrm{REL}} i}{\mathrm{REL}} = \frac{26 - 15}{5} = 2.2 \tag{13}$$

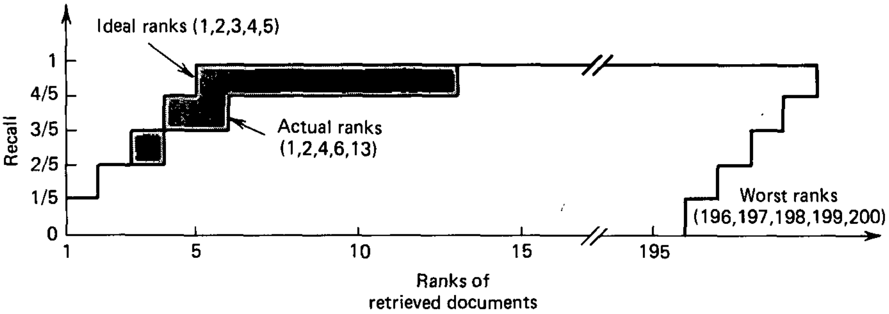


Figure 5-10 Construction of normalized recall measure.

where REL is the number of relevant documents and RANK_i represents the document ranks of the relevant items.

The values for expression (13) range from 0 for the case of perfect retrieval to N − REL for the worst possible case. That is, if the REL relevant documents are ranked 1, 2, 3, . . . , REL, then the value of expression (13) is equal to 0. On the other hand, if the relevant document ranks are N − REL + 1, N − REL + 2, . . . , N where N is the number of documents in the collection, then the value of expression (13) is equal to N − REL. Hence expression (13) can be normalized by dividing by N − REL. Finally, subtraction from 1 ensures that the measure equals 1 for the best case and 0 for the worst instead of vice versa. The resulting measure, known as the normalized recall, is then given by

$$\mathrm{RECALL}_{\mathrm{norm}} = 1 - \frac{\sum_{i=1}^{\mathrm{REL}} \mathrm{RANK}_i - \sum_{i=1}^{\mathrm{REL}} i}{\mathrm{REL}(N - \mathrm{REL})} \tag{14}$$

For the case shown in Fig. 5-10, $\mathrm{RECALL}_{\mathrm{norm}}$ equals 0.989. This reflects the fact that the ranks of the relevant documents deviate very little from the ideal case.

The tinted area in Fig. 5-10 reflects the number of nonrelevant documents that have to be retrieved in order to reach a recall value of 1. Since the latter measure is akin to the fallout, the normalized recall value of expression (14) may in fact be shown to be equivalent to the area under the recall-fallout curve of Fig. 5-6b [42].

An equivalent development for the computation of the normalized area between actual and ideal precision curves leads to a normalized precision measure defined as

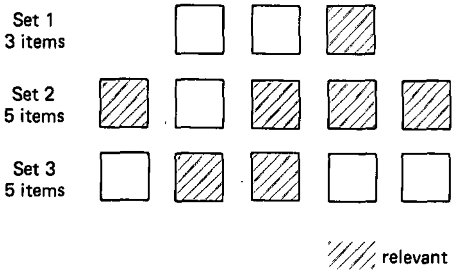
$$\mathrm{PRECISION}_{\mathrm{norm}} = 1 - \frac{\sum_{i=1}^{\mathrm{REL}} \log \mathrm{RANK}_i - \sum_{i=1}^{\mathrm{REL}} \log i}{\log N!/(N - \mathrm{REL})!\mathrm{REL}!} \tag{15}$$

Just like ordinary recall and precision, the normalized recall is sensitive to the rank assigned to the last relevant document in the retrieval order, and the normalized precision is sensitive to the rank of the first relevant document in the retrieval order. The normalized measures constitute a summary of the full recall-precision curve; as such they cannot be computed for a single retrieval point corresponding to a single recall-precision pair.

A number of additional single-valued measures also use the differences between the actual ranks of the relevant items retrieved, and either the ideal ranks where all the relevant items are retrieved ahead of any nonrelevant item, or a random system where the relevant items are randomly sprinkled among the nonrelevant. The expected search length [52] and the sliding ratio [53] are two measures of this kind.

Consider first the *expected search length*. Here one assumes that documents are presented to the user in a weakly ordered sequence following an information search—for example, all documents exhibiting NMATCH matching terms with the query would be retrieved before the set with NMATCH − 1 matching terms, and those in turn would precede the set with NMATCH − 2 matching terms, and so on. The *search length* may then be defined as the average number of nonrelevant items that must be scanned by the user before the total number of wanted items is reached.

Consider as an example the case of Fig. 5-11 which includes 3 items in set 1



(a)

Number of relevant wanted	Number of sets to be searched	Search length	Average search length
1	1	0, 1, or 2	1/3·0 + 1/3·1 + 1/3·2 = 1
6	3	3, 4, 5, or 6	4/10·3 + 3/10·4 + 2/10·5 + 1/10·6 = 4

(b)

Figure 5-11 Average search length illustration. (a) Partly ordered retrieval output. (b) Average search length computation. (*Adapted from reference 52.*)

and 5 items in each of sets 2 and 3. Set 1 is retrieved before set 2 and set 2 is retrieved before set 3. Within each of the sets the documents are not ranked or ordered. That is, there is no obvious way to order the documents having NMATCH terms in common with the query. The average search length necessary to retrieve 6 relevant items is 4. That is, the searcher will have to examine 4 nonrelevant documents on the average in order to find 6 relevant ones. In particular, since only one relevant item is retrievable from set 1, and 4 more from set 2, it is always necessary to look into set 3 for one additional relevant item to make up the total of 6. The first relevant item in set 3 may be located in position 1, producing a total search length of 3; or in positions 2, 3, or 4, producing search lengths of 4, 5, or 6, respectively. Of the 10 possible ways of distributing two relevant items among 5 positions in set 3, 4 have a relevant item in position 1, 3 more in position 2, 2 more in position 3, and 1 in position 4. This results in the search length computation shown in Fig. 5-11.

Now consider a QUERY and let PREVNREL be the number of nonrelevant documents in all sets preceding the one where the search terminates. If there are REL relevant items in the final set, and if they are put at equal intervals among the NREL nonrelevant documents in that set, then REL + 1 subsequences of nonrelevant documents will normally be created containing NREL/REL + 1 nonrelevant documents each. If one assumes that the request is satisfied at the NUMth relevant item on the last level, the expected search length EXP for QUERY will be

$$\text{EXP(QUERY)} = \text{PREVNREL} + \frac{\text{NREL} \cdot \text{NUM}}{\text{REL} + 1} \tag{16}$$

The expected random search length is obtained by scattering the ALLREL documents relevant to QUERY randomly through the IRREL irrelevant items. It is defined as DESIRED · IRREL/(ALLREL + 1), where DESIRED is the total desired number of relevant. A useful measure is specified as the improvement obtained by the actual case EXP over the random case REXP [51] as follows:

$$\begin{aligned}\text{EXP(QUERY) REDUCTION} &= \frac{\text{REXP(QUERY)} - \text{EXP(QUERY)}}{\text{REXP(QUERY)}} \\ &= 1 - \frac{\text{PREVNREL} + \dfrac{\text{NREL} \cdot \text{NUM}}{\text{REL} + 1}}{\dfrac{\text{DESIRED} \cdot \text{IRREL}}{\text{ALLREL} + 1}} \end{aligned} \tag{17}$$

The *sliding ratio* measure is based on the comparison of two ranked lists of items. One list is the output of an actual retrieval system, and the other represents an ideal system in which the items are ranked in decreasing relevance order [53]. This model is more complex than the ones previously described because it allows the assignment of numeric relevance weights to the documents.

Table 5-5 Sample Computation for Sliding Ratio Measure

Retrieval rank	1	2	3	4	5	
Document number	3	4	5	1	2	Actual system output
Relevance weight $\text{WEIGHT}_i^{\text{REAL}}$	10	0	8	5	2	
$\text{SWEIGHT REAL(NUM)} = \sum_{i=1}^{\text{NUM}} \text{WEIGHT}_i^{\text{REAL}}$	10	10	18	23	25	
Document number	3	5	1	2	4	Ideal system output
Relevance weight $\text{WEIGHT}_i^{\text{IDEAL}}$	10	8	5	2	0	
$\text{SWEIGHT IDEAL(NUM)} = \sum_{i=1}^{\text{NUM}} \text{WEIGHT}_i^{\text{IDEAL}}$	10	18	23	25	25	
$\text{SLIDE(NUM)} = \dfrac{\text{SWEIGHT REAL(NUM)}}{\text{SWEIGHT IDEAL(NUM)}}$	1	0.55	0.78	0.92	1	

Adapted from reference 53.

These replace the usual binary relevance assessments. In an ideal retrieval situation, the documents would be ranked in decreasing order of their relevance weights rather than by a simple precedence of the relevant set ahead of the set of nonrelevant items. Ties in rank and in relevance values between items can be eliminated, for example by assigning to all items in each set the average rank of the items in the set.

The sliding ratio measure SLIDE(NUM) for retrieval cutoff SLIDE may be defined as

$$\text{SLIDE(NUM)} = \frac{\text{SWEIGHT REAL(NUM)}}{\text{SWEIGHT IDEAL(NUM)}} \tag{18}$$

where SWEIGHT REAL(NUM) and SWEIGHT IDEAL(NUM) are the sum of the relevance weights of all items retrieved up to rank NUM in the actual and ideal systems, respectively. A sample calculation is shown in Table 5-5 for five documents which do not exhibit any ties in rank. The value of SLIDE(NUM) at any particular value of NUM shows the ability of the actual system to approximate the retrieval capability of the ideal system. In the limit, as NUM approaches the total number of documents in the collection, WEIGHT REAL(NUM) becomes equal to WEIGHT IDEAL(NUM) and SLIDE(NUM) becomes equal to 1.

The sliding ratio can of course also be used for systems where the relevance weights are restricted to 1 for relevant items and 0 for the nonrelevant. In that case, the ratio approximates the normalized recall and the expected search length.

**D Utility Measure

A property shared by all the measures described in the preceding sections is the fact that only system effectiveness is taken into account. The cost or value of a particular retrieval action has not been considered. Assuming that cost and/or value parameters are available, it is possible to devise retrieval evaluation strategies based on an extension of the standard contingency table as shown in Table 5-6 [54,55]. The data in Table 5-6 show the usual four-way split of the document collection into the number of items relevant and retrieved RETREL, the number of items retrieved and nonrelevant RETNREL, the number of items not retrieved but relevant NRETREL and the number of items not retrieved and not relevant NRETNREL. A value of VALUE_1 is assigned to each relevant item that is retrieved, and VALUE_2 is assigned to each nonrelevant item that is rejected. Similarly, costs of COST_1 and COST_2 are associated with nonrelevant items retrieved and with relevant items that are missed.

If the value of the similarity measure between a document DOC and a query QUERY can be expressed as a variable VAR = FUNC(QUERY,DOC), then the *utility* of a given relevant document set DOCSET with respect to some query QUERY at retrieval threshold VAR = THRESHOLD can be expressed as

$$\begin{aligned}\text{UTIL(DOCSET,QUERY, THRESHOLD)} &= \text{VALUE}_1 \cdot \text{RETREL} \\ &\quad - \text{COST}_1 \cdot \text{RETNREL} - \text{COST}_2 \cdot \text{NRETREL} \\ &\quad + \text{VALUE}_2 \cdot \text{NRETNREL}\end{aligned} \tag{19}$$

or alternatively as

$$\begin{aligned}&\text{UTIL(DOCSET,QUERY,THRESHOLD)} = \\ &\quad \text{VALUE}_1 \cdot \text{N Prob}\{\text{DOC is relevant and VAR} \geqslant \text{THRESHOLD}\} \\ &\quad - \text{COST}_1 \cdot \text{N Prob}\{\text{DOC is not relevant and VAR} \geqslant \text{THRESHOLD}\} \\ &\quad - \text{COST}_2 \cdot \text{N Prob}\{\text{DOC is relevant and VAR} < \text{THRESHOLD}\} \\ &\quad + \text{VALUE}_2 \cdot \text{N Prob}\{\text{DOC is not relevant and VAR} < \text{THRESHOLD}\}\end{aligned} \tag{20}$$

where N is the total number of documents in the system. Expression (20) can be transformed using the probability density functions $\text{FUNC}_1(\text{VAR})$ and $\text{FUNC}_2(\text{VAR})$ previously introduced in Fig. 5-7. In fact, the area under the density curves to the right of a given threshold represent the probabilities that variable VAR has a value greater than the threshold, given that the documents are relevant and nonrelevant, respectively. By substitution of integrals into expression (20) a useful retrieval threshold is obtained for which the utility of the system is positive.

Table 5-6 Contingency Table with Cost and Value Parameters

	Relevant	Nonrelevant	
Retrieved	v_1 (RETREL)	c_1 (RETNREL)	RETREL + RETNREL
Not retrieved	c_2 (NRETREL)	v_2 (NRETNREL)	NRETREL + NRETNREL
	RETREL + NRETREL	RETNREL + NRETNREL	N

v_1 = VALUE_1; v_2 = VALUE_2
c_1 = COST_1; c_2 = COST_2

A substantial literature exists relating to the use of the utility measure for retrieval system evaluation [54–57]. However, until simple methods become available for estimating the cost and value parameters for the individual documents in a collection, the theoretical appeal of this method may outweigh its practical usefulness.

4 EVALUATION OF SYSTEM COST AND EFFICIENCY

A System Tradeoffs

The art of efficiency analysis is not as far advanced as the analysis of system effectiveness. This is because accurate cost data in terms of time, effort, and money spent are difficult to obtain, and because the value of improved information services and the benefits derivable from them is impossible to ascertain in most environments. Furthermore, when identifying information system costs, invariably one is forced to look at noncomparable situations. The cost differences between two systems, such as an automated and a manual one, may not accurately reflect the value of either system. The automated system might, for example, be used for purposes other than information storage and retrieval, or it might be usable on a 24-hour per day basis, whereas a manual system might not. Thus an efficiency evaluation involves a great many intangible factors which may hamper a concrete analysis and render the results unreliable or meaningless.

Nevertheless, it is necessary to consider the cost analysis question. Information systems are not likely to be constructed or installed without some attempt at evaluating their potential efficiency. It is customary to distinguish between *cost-effectiveness* analysis and *cost-benefit* analysis. The former is designed to find the least expensive means for carrying out a given set of operations or to obtain the maximum value from a given expenditure. Cost-benefit analysis requires a systematic comparison between the costs of individual operations and the benefits derivable from them [58–60].

The costs of a system can be divided into the *initial* development *costs* necessary for design, testing, and evaluation; the *operating costs* which are variable and depend on the tasks performed, the personnel used, and the amount of equipment required; and finally, the *fixed costs* for rent, taxes, and other standard items. The benefits obtainable from the information system may be related to decreased costs or increased productivity. Cost savings are difficult to document when manual operations are replaced by automatic ones. It is even harder to measure the benefits of sophisticated information systems which may consist of improved decision making capabilities, increased productivity, stimulation of research capacity, and the like, and the value of these somewhat serendipitous factors is normally impossible to ascertain.

In an information retrieval situation in which the volume of operations—such as number of documents, size and cost of the documents, and average number of queries—is given, the basic alternatives and system tradeoffs relate to the input and document indexing operations on the one hand, and to the information search and output transactions on the other. A particular performance criterion—for example, a given precision level—can normally be attained in many different ways, each of them involving different cost levels. Thus, precision may be raised by using a highly specific indexing vocabulary requiring high indexer proficiency and large indexing costs. Alternatively, the indexing may be performed more casually, but the output might be screened by trained subject experts before presentation to the users, thereby decreasing indexing costs but lengthening search time. Finally, the burden might be shifted to the user, by having customers conduct an interactive search and letting them rephrase the query formulation in the hope of generating better output.

In some cases, it is possible to obtain quantifiable information which relates various system alternatives to the effectiveness or quality of the output product. The following relationships may be cited as examples [7,58,59]:

1 Collection coverage versus expected number of retrievals; normally, a very small proportion of items accounts for a large proportion of all relevant items retrieved; the cost of adding to the collection a large number of the less productive items may thus be difficult to justify in terms of improvements in the output product.

2 Indexing time versus search effectiveness; there is a direct relation between indexing time and indexing exhaustivity and the corresponding expected recall; unfortunately, at high recall, the required indexing time increases much more rapidly than the recall performance, so that diminishing returns set in when the indexing time or exhaustivity exceed a given limit.

3 Specificity of the indexing language and recall-precision balance; normally, a more specific indexing language costs more to develop and produces better precision but may cause losses in recall; obviously, the desirable level of precision and thus the importance of language specificity varies with collection size, high precision being most crucial for very large collections.

4 Equipment complexity versus processing limitations; in general, more sophisticated equipment can produce a greater variety of output products—for example, ranked output consisting of document abstracts, instead of unranked document numbers or titles; on the other hand, more sophisticated processing devices cost more to acquire and to operate and put a greater burden on the system operators, and sometimes on the users.

Even if the various system alternatives are quantifiable in a reliable way, it may be difficult to reach operational decisions because the large fixed costs associated with an implementation may not be easily recoverable by instituting fees for service provided. Until agreement is reached concerning the value and benefits of information services, a cost analysis is not likely to produce the answers by management.

**B Cost Analysis

There exist two basically distinct approaches to the analysis of the costs of an information system. The first one consists in carefully analyzing the various steps included in an actual processing chain, and in performing direct measurements of the various quantities which enter into the cost picture for a given operation environment. The second consists in generating an abstract model of the system being investigated, and in ascertaining system efficiency by carrying out appropriate simulation studies. In either case, all actual as well as hidden costs ought to be taken into account, including development costs, operating costs, and fixed costs.

Consider first the approach which starts with actual system measurements. As might be expected, volumes of published cost figures may be found in the literature presenting a wide array of measure data [61–64]. In general, the published data are unrelatable to each other, because of differences in the respective environments and in the assumptions made when performing the measurements. However, the published values do make it possible to obtain an idea of the relative expense arising from various processing steps, and occasionally the absolute magnitude of some item—for example, a stated manual cataloging cost of over $10 per item obtained for five large university libraries in 1969—may in itself furnish cause for concern. A unit cost figure is, however, not as useful as an indicator of system efficiency as a calculated cost related to some effectiveness measurement, such as, for example, the cost per citation retrieved by the system or, better still, the cost per relevant citation retrieved [65,66].

A typical efficiency analysis based on initial measurements of *costs, time, and volume* of operations would start with a formal system description, including a specification of the interrelationships between processes, and the generation of basic parameter values relating file sizes, input and output rates, and other operating characteristics. Several functional models of this type exist for libraries and information centers, normally including information acquisition, data encoding or indexing, storage organization, query preparation, information search, output operations, and in some instances also user appraisal and feedback operations leading to query reformulations [67,68].

The statement of time and cost data for equipment, personnel, materials, and procedures, and a specification of the interdependencies between various system parameters then lead to the generation of *performance measurements* which relate system performance to user requirements. The user requirements may be specified in terms of output volume, response time, recall and precision requirements, and the like. If a computer program is used for the computation of the functions, it may be possible to perform measurements for various assumed levels of the parameter values—for example, using different input rates for new materials, and different monthly search volumes. This leads to the generation of different figures of merit for different assumed operating conditions, and to decisions concerning a possible expansion of services or to transformations in the current operating practices.

The cost-time-volume model is valuable in situations where exact system specifications and parameter values are available. In practice these values are usually not available, and when they are, a computation of performance measurements closely tied to current operating characteristics is usually not needed, since the system operations may already be well in hand. In such cases, a more ambitious *system simulation* might be undertaken leading to a theoretical analysis of new concepts and ideas, including system growth studies, error and reliability studies, and comparative evaluations of new system configurations [69,70].

Consider now the specification of actual cost functions. A variety of formulations have been used for this purpose, including some based on cost-benefit comparisons and on the dual use of both efficiency as well as effectiveness criteria [71–74]. One possibility consists, for example, in assuming that costs may be subdivided into four types, including initial costs for development, operating costs for personnel and materials, fixed costs such as rent and taxes, and finally operating returns derived from the sale of products [75]. If the returns are disregarded, development costs are broken down into designing, testing, operating, and reporting, and operating costs are separated into three parts, including clerical, machine, and technical and professional costs, the resulting cost function may be expressed as

$$\begin{aligned}\mathrm{COST} &= \mathrm{C(TIME,UNITS)}\\ &= \sum_{i=1}^{\mathrm{FIXED}} \mathrm{FIX}_i\mathrm{(TIME)} + \mathrm{DVLP(TIME)} + \mathrm{OPER(TIME,UNITS)} \qquad (21)\end{aligned}$$

where TIME = unit time
UNITS = the number of unit operations to be considered
FIX_i = ith subdivision of the fixed costs in dollars per unit of time
FIXED = number of subdivisions of fixed cost
DVLP = development cost in dollars per unit time, amortized as current cost
OPER = operating cost in dollars per unit time per unit operation
C(TIME,UNITS) = general cost function varying with time and number of operations

The various factors of equation (21) can be further broken down into individual components. Thus, the operating costs can be expressed as

$$\begin{aligned}\mathrm{OPER(TIME,UNITS)} = \mathrm{UNITS}\Bigg[&\sum_{j=1}^{\mathrm{CLERICAL}} \mathrm{CLER}_j\mathrm{(TIME)TIM}_{\mathrm{CLRK},j}\\ &+ \sum_{k=1}^{\mathrm{MACH}} \mathrm{MCH}_k\mathrm{(TIME)TIM}_{\mathrm{MAC},k}\\ &+ \sum_{q=1}^{\mathrm{TECH}} \mathrm{TCH}_q\mathrm{(TIME)TIM}_{\mathrm{TNCH},q}\Bigg] \qquad (22)\end{aligned}$$

where CLER_j = jth subdivision of the clerical costs in dollars per unit, and time per unit operation (for example, the salary rate of the typists in a certain category)
CLERICAL = number of subdivisions of clerical cost

MCH_k = kth subdivision of machine costs
MACH = number of subdivisions of machine costs
TCH_q = qth subdivision of technical cost
TECH = number of subdivisions of technical cost
$TIM_{CLRK,j}$, $TIM_{MAC,k}$, and $TIM_{TNCH,q}$ = increments of time used for clerical, machine, and technical operations, respectively

Relationships may also be used between some of the cost components in order to obtain optimum allocations for some of the subunits. Thus the operating cost of individual operations often decreases as the response time is allowed to increase; at the same time the user's "delay cost" increases with time. An optimal response time must then exist, where the operating costs are no longer maximal (as they would be if instantaneous responses were demanded), while the output delay cost is still reasonable [76].

Analogously, search cost could be compared with collection size, where the cost first decreases with increasing file size to some optimal value as the proportion of items that can be stored in internal machine memory increases. Then as the file size increases further, all added items are stored in external, slow access memory, thereby increasing access and search costs. A typical cost curve of this type derivable by appropriate efficiency evaluation methods is shown in Fig. 5-12 [77].

A final cost model to be mentioned is based on a comparison between costs and benefits [78]. Two parameters are defined first, known, respectively, as the benefit to cost ratio BENEFIT/COST, and the PROFIT, equal to BENEFIT − COST. One assumes that the cost COST varies with respect to three parameters: the number of documents $NDOC_F$ in the file, the number of SEARCHES conducted per year, and the number of documents DOCPERSEARCH retrieved per search. All other costs are assumed fixed. A typical annual cost function might then be

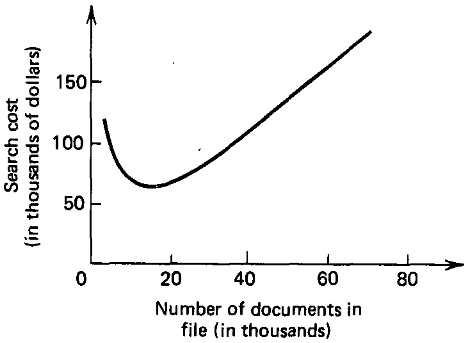


Figure 5-12 Typical cost curve reflecting search cost. (*Adapted from reference 77.*)

$$\text{COST} = \text{FIXED} + \text{MARGE} \cdot \text{NDOC}_F + (\text{FIXEDSEARCH} + \text{VARYSEARCH} \cdot \text{DOCPERSEARCH}) \cdot \text{SEARCHES} \quad (23)$$

where FIXED = fixed costs
MARGE = marginal costs of storing an additional item
FIXEDSEARCH = fixed search cost
VARYSEARCH = search cost that varies as a function of the number of output items examined

If one postulates that the benefit derived by a user from a search varies as the fraction of relevant items identified by the search from 0 to 1, then the benefit from a given imperfect search is BENEFIT · VALUE, where VALUE is the fractional user benefit derived from a given search, and BENEFIT is the maximal obtainable user benefit (in dollars). For NUM searches, the yearly benefit is then BENEFIT_T = BENEFIT · VALUE · NUM. In these circumstances the annual net benefit $\text{PROFIT}_T = \text{BENEFIT}_T - \text{COST}_T$ will be

$$\begin{aligned}\text{PROFIT}_T = {} & (\text{BENEFIT} \cdot \text{VALUE} - \text{FIXEDSEARCH} \\ & - \text{VARYSEARCH} \cdot \text{DOCPERSEARCH})\text{SEARCHES} \\ & - \text{MARGE} \cdot \text{NDOC} - \text{FIXED}\end{aligned} \quad (24)$$

The simple net benefit model of equation (24) serves only as an approximation to a much more complicated real situation. However when reasonably accurate sample values are used for the various parameters, useful indications may be obtainable from efficiency and effectiveness evaluations reflecting the behavior of the system in the real-life environments that are being investigated.

5 SUMMARY

It should be clear that the most dominant form of evaluation remains the precision-recall curve defined early in this chapter. However, it is also obvious that a great deal of effort has been invested to develop other means of evaluating information retrieval systems. Some of these are reasonably practical even if seldom used, such as fallout, and others are theoretically interesting but almost never used, such as Swets' E.

Difficulties arise in acquiring much of the data required for evaluation. Precision and recall seem to present relatively few conceptual problems, although both depend on the availability of objective relevance assessments of documents with respect to queries. Precision and recall are also readily interpretable in terms of the actual performance of the system, and they may be relatable to cost evaluation by using composite measures such as, for example, the cost per relevant document retrieved.

Since many decisions eventually revolve around system cost, it would be convenient if one could easily measure either the cost or the benefit of a system. Unfortunately, costs and values vary from environment to environment, and the expense associated with a particular system function or operation is often impossible to isolate from the surrounding context. Thus, while some cost evaluation methods have been presented, it must be recognized that the collection of cost data is very difficult. Furthermore, the comparison of costs pertaining to different environments is particularly dangerous.

In the remainder of this book, the evaluation results are based on recall and precision, because these measures remain the standard. In operational retrieval environments, many other factors may, however, prove more important than recall and precision, especially to uninitiated users of the system. Whereas a small improvement in either recall or precision may be completely invisible to the average user, human factor considerations such as ease of use and training required for query submission, output formats, console noise, system reliability, and response time may take on overwhelming importance.

The evaluation measures examined in this chapter make it possible to distinguish well-designed methods from other less effective ones. In the end, acceptable retrieval systems must be easy to use, reliable, effective and inexpensive. All system users look forward to the day when such systems will actually become available for use.

REFERENCES

[1] D.W. King and E.C. Bryant, The Evaluation of Information Services and Products, Information Resources Press, Washington, 1971.

[2] A. Kent, O.E. Taulbee, J. Belzer, and G.D. Goldstein, editors, Electronic Handling of Information: Testing and Evaluation, Thompson Book Co., Washington, 1967.

[3] S. Treu, Testing and Evaluation—Literature Review, in Electronic Handling of Information: Testing and Evaluation, A. Kent, O.E. Taulbee, J. Belzer, and G.D. Goldstein, editors, Thompson Book Co., Washington, 1967, pp. 71–88.

[4] F.W. Lancaster, Evaluating the Effectiveness of Information Retrieval Systems, in Information Retrieval Systems—Characteristics, Testing and Evaluation, 2nd Edition, Chapter 9, John Wiley and Sons, New York, 1979.

[5] C.J. Wessel, Criteria for Evaluating Technical Library Effectiveness, Aslib Proceedings, Vol. 20, No. 11, November 1968, pp. 455–481.

[6] H. Bornstein, A Paradigm for a Retrieval Effectiveness Experiment, American Documentation, Vol. 12, No. 4, October 1961, pp. 254–481.

[7] F.W. Lancaster and W.D. Climenson, Evaluating the Economic Efficiency of a Document Retrieval System, Journal of Documentation, Vol. 24, No. 1, March 1968, pp. 16–40

[8] C.P. Bourne, Review of the Criteria and Techniques Used or Suggested for the Evaluation of Reference Retrieval Systems, Report, Stanford Research Institute, Menlo Park, California, September 1964.

[9] M.B. Snyder, A.W. Schumacher, S.E. Mayer, and M.D. Havron, Methodology for Test and Evaluation of Document Retrieval Systems: A Critical Review and Recommendations, Report to the National Science Foundation, Human Sciences Research Inc., McLean, Virginia, January 1966.

[10] D.W. King and E.C. Bryant, A Diagnostic Model for Evaluating Retrospective

Search Systems, Information Storage and Retrieval, Vol. 6, No. 3, July 1970, pp. 261–272.

[11] T. Saracevic, Linking Research and Teaching, American Documentation, Vol. 19, No. 4. October 1968, pp. 398–403.

[12] F.W. Lancaster, The Functions of Information Retrieval Systems, in Information Retrieval Systems—Characteristics, Testing and Evaluation, 2nd Edition, Chapter 1, John Wiley and Sons, New York, 1979.

[13] C.W. Cleverdon, J. Mills, and E.M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 1—Design, Aslib Cranfield Research Project, Cranfield, England, 1966.

[14] F.W. Lancaster, Criteria by which Information Retrieval Systems May Be Evaluated, in Information Retrieval Systems—Characteristics, Testing and Evaluation, 2nd Edition, Chapter 8, John Wiley and Sons, New York, 1979.

[15] C.A. Cuadra and R.V. Katter, Experimental Studies of Relevance Judgments Report TM-3520, Final Report, Vol. 1, System Development Corporation, Santa Monica, California, June 1967.

[16] T. Saracevic, Relevance: A Review of and a Framework for the Thinking on the Notion in Information Science, Journal of the ASIS, Vol. 26, No. 6, November–December 1975, pp. 321–343.

[17] W.S. Cooper, A Definition of Relevance for Information Retrieval, Information Storage and Retrieval, Vol. 7, No. 1, June 1971, pp. 19–37.

[18] W. Goffman, On Relevance as a Measure, Information Storage and Retrieval, Vol. 2, No. 3, December 1964, pp. 201–203.

[19] W. Goffman and V.A. Newill, Methodology for Test and Evaluation of Information Retrieval Systems, Comparative Systems Laboratory, Report CSL: TR-2, Western Reserve University, Cleveland, Ohio, July 1964.

[20] D.A. Kemp, Relevance, Pertinence and Information System Development, Information Storage and Retrieval, Vol. 10, 1974, pp. 37–47.

[21] S.E. Robertson, The Probabilistic Character of Relevance, Information Processing and Management, Vol. 13, No. 4, 1977, pp. 247–251.

[22] M.E. Maron and J.L. Kuhns, On Relevance, Probabilistic Indexing and Information Retrieval, Journal of the ACM, Vol. 7, No. 3, July 1960, pp. 216–244.

[23] C.W. Cleverdon and J. Mills, The Testing of Index Language Devices, Aslib Proceedings, Vol. 15, No. 4, April 1963, pp. 106–130.

[24] G. Salton, editor, The SMART Retrieval System—Experiments in Automatic Document Processing, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971, Part 3.

[25] C.J. van Rijsbergen, Information Retrieval, 2nd Edition, Chapter 7, Butterworths, London, 1979.

[26] H. Gilbert and K. Sparck Jones, Statistical Bases of Relevance Assessments for the Ideal Information Retrieval Test Collection, Computer Laboratory, University of Cambridge, BL R and D Report 5481, Cambridge, England, March 1979.

[27] E.M. Keen, Evaluation Parameters, in The SMART Retrieval System—Experiments in Automatic Document Processing, G. Salton, editor, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971, Chapter 5.

[28] R.A. Fairthorne, Basic Parameters of Retrieval Tests, Proceedings of 1964 Annual Meeting of the American Documentation Institute, Spartan Books, Washington, 1964, pp. 343–347.

[29] D. Williamson, R. Williamson, and M.E. Lesk, The Cornell Implementation of the

SMART System, in The SMART Retrieval System—Experiments in Automatic Document Processing, G. Salton, editor, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971, Chapter 2.

[30] F.W. Lancaster, Evaluating the Performance of a Large Operating Information Retrieval System, in Electronic Handling of Information: Testing and Evaluation, A. Kent, O.E. Taulbee, J. Belzer, and G.D. Goldstein, editors, Thompson Book Co., Washington, 1967, pp. 199–216.

[31] G. Salton, The "Generality" Effect and the Retrieval Evaluation for Large Collections, Journal of the ASIS, Vol. 23, No. 1, January–February 1972, pp. 11–22.

[32] M. Taube, A Note on the Pseudomathematics of Relevance, American Documentation, Vol. 16, No. 2, April 1965, pp. 69–72.

[33] M.E. Lesk and G. Salton, Relevance Assessments and Retrieval System Evaluation, Information Storage and Retrieval, Vol. 4, No. 4, December 1968, pp. 343–359.

[34] V.E. Giuliano and P.E. Jones, Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems, Report No. ESD-TR-66-405, Arthur D. Little, Inc., Cambridge, August 1966.

[35] R.A. Fairthorne, Implications of Test Procedures, in Information Retrieval in Action, Western Reserve University Press, Cleveland, Ohio, 1963, pp. 109–113.

[36] S.E. Robertson, The Parametric Description of Retrieval Tests, Part I: The Basic Parameters, Journal of Documentation, Vol. 25, No. 1, March 1969, pp. 1–27.

[37] C.W. Cleverdon, Progress in Documentation: Evaluation Tests of Information Retrieval Systems, Journal of Documentation, Vol. 26, No. 1, March 1970, pp. 55–67.

[38] T. Saracevic, An Inquiry into Testing of Information Retrieval Systems, Part I: Objectives, Methodology, Design and Controls, Comparative Systems Laboratory, Report No. CSL: TR-FINAL-1, Case Western Reserve University, Cleveland, Ohio, 1968.

[39] G.F. Romerio and L. Cavara, Assessment Studies of Documentation Systems, Information Storage and Retrieval, Vol. 4, No. 3, August 1968, pp. 309–325.

[40] J.A. Swets, Effectiveness of Information Retrieval Methods, American Documentation, Vol. 20, No. 1, January 1969, pp. 72–89.

[41] B.C. Brookes, The Measure of Information Retrieval Effectiveness Proposed by Swets, Journal of Documentation, Vol. 24, No. 1, March 1968, pp. 41–54.

[42] S.E. Robertson, The Parametric Description of Retrieval Tests, Part 1: The Basic Parameters, Journal of Documentation, Vol. 25, No. 1, March 1969, pp. 1–27; Part 2: Overall Measures, Journal of Documentation, Vol. 25, No. 2, June 1969, pp. 93–107.

[43] M.H. Heine, Design Equations for Retrieval Systems Based on the Swets Model, Journal of the ASIS, Vol. 25, No. 3, May–June 1974, pp. 183–198.

[44] R.R.V. Wiederkehr, Search Characteristics Curves, in Evaluation of Document Retrieval Systems: Literature Perspective, Measurement, Technical Papers, Westat Research Report PB 182710, Bethesda, Maryland, December 1968.

[45] D.W. King and E.C. Bryant, The Evaluation of Information Services and Products, Information Resources Press, Washington, 1971, Chapter 9.

[46] A.R. Meetham, Communication Theory and the Evaluation of Information Retrieval Systems, Information Storage and Retrieval, Vol. 5, No. 5, October 1969, pp. 129–134.

[47] R.H. Shumway, Contingency Tables in Information Retrieval: An Information Theoretic Analysis, in Evaluation of Document Retrieval Systems: Literature Per-

spective, Measurement, Technical Papers, Westat Research Report PB 182710, Bethesda, Maryland, December 1968.

[48] M. Guazzo, Retrieval Performance and Information Theory, Information Processing and Management, Vol. 13, No. 3, 1977, pp. 155–165.

[49] A.E. Cawkell, A Measure of Efficiency Factor—Communication Theory Applied to Document Selection Systems, Information Processing and Management, Vol. 11, No. 8–12, 1975, pp. 243–248.

[50] J.J. Rocchio, Jr., Document Retrieval Systems—Optimization and Evaluation, Harvard University Doctoral Thesis, Report No. ISR-10 to the National Science Foundation, Harvard Computation Laboratory, March 1966.

[51] G. Salton, Automatic Information Organization and Retrieval, McGraw-Hill Book Co., New York, 1968, Chapter 8.

[52] W.S. Cooper, Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems, American Documentation, Vol. 19, No. 1, January 1968, pp. 30–41.

[53] S.M. Pollock, Measures for the Comparison of Information Retrieval Systems, American Documentation, Vol. 19, No. 4, October 1968, pp. 387–397.

[54] W.S. Cooper, On Selecting a Measure of Retrieval Effectiveness, Part I: The Subjective Philosophy of Evaluation; Part II: Implementation of the Philosophy, Journal of the ASIS, Vol. 24, 1973, pp. 87–100, 413–424.

[55] D.H. Kraft and A. Bookstein, Evaluation of Information Retrieval Systems: A Decision Theory Approach, Journal of the ASIS, Vol. 29, No. 1, January 1978, pp. 31–40.

[56] W.S. Cooper, Indexing Documents by Gedanken Experimentation, Journal of the ASIS, Vol. 29, No. 3, May–June 1978, pp. 107–119.

[57] W.S. Cooper and M.E. Maron, Foundations of Probabilistic and Utility Theoretic Indexing, Journal of the ACM, Vol. 25, No. 1, January 1978, pp. 67–80.

[58] F.W. Lancaster, The Cost-Effectiveness Analysis of Information Retrieval and Dissemination Systems, Journal of the ASIS, Vol. 22, No. 1, January 1971, pp. 12–27.

[59] F.W. Lancaster, Cost-Effectiveness and Cost-Benefit Evaluation, in Information Retrieval Systems—Characteristics, Testing and Evaluation, 2nd Edition, John Wiley and Sons, New York, 1979, Chapter 16.

[60] A. Gilchrist, Cost-Effectiveness, Aslib Proceedings, Vol. 23, No. 9, September 1971, pp. 455–464.

[61] F. Alouche, N. Bely, R.C. Cros, J.C. Gardin, F. Levy, and J. Perreault, Economie Générale d'une Chaîne Documentaire Mecanisée, Gauthier Villars, Paris, 1967.

[62] K.W. Webb, W.C. Suhler, G.G. Heller, and S.P. Todd, Jr., Evaluation Models for Information Retrieval and Command and Control Systems (EMIR), IBM Corporation Report, Federal Systems Division, Washington, June 1964.

[63] A Time/Cost Study of Processing Books via Unit Orders and Blanket Orders, Five Associated University Libraries, Newsletter, Vol. 3, No. 4, July 1972.

[64] G. Williams, E.C. Bryant, R.R.V. Wiederkehr, V.E. Palmour, and C.J. Siehler, Library Cost Models: Owning versus Borrowing Serial Publications, Center for Research Libraries, Washington, November 1968.

[65] M.M. Cummings, Needs of the Health Sciences, in Electronic Handling of Information: Testing and Evaluation, A. Kent, O.E. Taulbee, J. Belzer and G.D. Goldstein, editors, Thompson Book Co., Washington, 1967.

[66] C.W. Cleverdon, The Methodology of Evaluation of Operational Information Re-

trieval Systems Based on the Test of Medlars, Cranfield Institute of Technology Report, Cranfield, England, 1968.

[67] C.P. Bourne, G.D. Peterson, B. Lefkowitz, and D. Ford, Requirements, Criteria and Measure of Performance of Information Storage and Retrieval Systems, Final Report to National Science Foundation, Report AD 270 942, Stanford Research Institute, December 1961.

[68] Arthur Andersen and Co., Research Study of Criteria and Procedures for Evaluating Scientific Information Retrieval Systems, Final Report to the National Science Foundation, New York, March 1962.

[69] C.P. Bourne and D.F. Ford, Cost Analysis and Simulation Procedures for the Evaluation of Large Information Systems, American Documentation, Vol. 15, No. 2, April 1964, pp. 142–149.

[70] N.R. Baker and R.E. Nance, The Use of Simulation in Studying Information Storage and Retrieval Systems, American Documentation, Vol. 19, No. 4, October 1968, pp. 363–370.

[71] M.D. Cooper, A Cost Model for Evaluating Information Retrieval Systems, Journal of the ASIS, Vol. 23, No. 5, September–October 1972, pp. 306–312.

[72] B.V. Tell, Auditing Procedures for Information Retrieval Systems, Proceedings 1965 FID Congress, Spartan Books, Washington, 1966, pp. 119–124.

[73] I.J. Good, The Decision-Theory Approach to the Evaluation of Information Retrieval System, Information Storage and Retrieval, Vol. 3, No. 2, April 1967, pp. 31–34.

[74] J. Martyn and B.C. Vickery, The Complexity of the Modelling of Information Systems, Journal of Documentation, Vol. 26, No. 3, September 1970, pp. 204–220.

[75] D.H. Rothenberg, An Efficiency Model and a Performance Function for an Information Retrieval System, Information Storage and Retrieval, Vol. 5, No. 3, October 1969, pp. 109–122.

[76] N.R. Keith, Jr., A General Evaluation Model for an Information Storage and Retrieval System, Journal of the ASIS, Vol. 21, No. 4, July–August 1970, pp. 237–239.

[77] E.C. Bryant, Modeling in Document Handling, in Electronic Handling of Information: Testing and Evaluation, A. Kent, O.E. Taulbee, J. Belzer, and G.D. Goldstein, editors, Thompson Book Co., Washington, 1967, pp. 163–173.

[78] R.R.V. Wiederkehr, A Net Benefit Model for Evaluating Elementary Document Retrieval Systems, in Evaluation of Document Retrieval Systems, Westat Research Report, PB 182710, Bethesda, Maryland, December 1968.

BIBLIOGRAPHIC REMARKS

Many materials covering the evaluation of information retrieval operations and systems appear in the report literature that may not be easy to obtain. Two of the best known of these reports cover the well-known Aslib-Cranfield study and the evaluation of the MEDLARS retrieval service:

C.W. Cleverdon, J. Mills, and E.M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 1—Design, Aslib-Cranfield Research Project, Cranfield, England, 1966.

F.W. Lancaster, Evaluation of the Medlars Demand Search Service, National Library of Medicine, Bethesda, Maryland, January 1968.

The following references contain discussions dealing with theoretical aspects of retrieval system evaluation:

T. Saracevic, Relevance-A Review of and a Framework for the Thinking on the Notion in Information Science, Journal of the American Society for Information Science, Vol. 26, No. 6, November–December 1975, pp. 321–343.

M.E. Maron, editor, Theory and Foundations of Information Retrieval, Drexel Library Quarterly, Vol. 14, No. 2, April 1978.

C.J. van Rijsbergen, Information Retrieval, 2nd Edition, Chapter 7, Butterworths, London, 1979.

The following texts all deal extensively with various aspects of retrieval system evaluation:

F.W. Lancaster, Information Retrieval Systems—Characteristics, Testing and Evaluation, 2nd Edition, John Wiley and Sons, New York, 1979.

A. Kent, O.E. Taulbee, J. Belzer, and G.D. Goldstein, editors, Electronic Handling of Information: Testing and Evaluation, Thompson Book Company, Washington, DC., 1967.

D.W. King and E.C. Bryant, The Evaluation of Information Services and Products, Information Resources Press, Washington, DC., 1971.

EXERCISES

5-1 Consider a retrieval system capable of presenting the output items to the user population in a ranked sequence in decreasing order of presumed usefulness. Consider two particular queries each having 10 relevant documents in a collection. The ranks of the relevant documents for query 1 are 1, 3, 5, . . . ,19, and for query 2 the ranks are 2, 4, 6, . . . ,20.

- **a** Prepare recall-precision tables and graphs for the two queries similar to those shown in Fig. 5-2 for a sample query.
- **b** What is the most obvious difference in the evaluation results obtained for the two queries? Do you expect this difference to affect a systems evaluation in which results are averaged over several queries? Why?
- **c** Prepare recall-precision tables and graphs showing recall-level and document-level averages for the recall and precision results obtained for the two queries.

5-2 The fallout evaluation measure has been called superior to precision for a number of reasons. What are they? Under what circumstances would you prefer to deal with a recall-fallout evaluation instead of a recall-precision output?

5-3 Prepare probability density output and operating characteristic curves similar to those shown in Figs. 5-7 and 5-8 reflecting the performance of the set of relevant items and the set of nonrelevant items, respectively, with respect to some query for the following cases:

- **a** All relevant items are retrieved ahead of all nonrelevant ones.
- **b** All nonrelevant items are retrieved ahead of all relevant ones.
- **c** The relevant items are randomly sprinkled among the nonrelevant ones.
- **d** The retrieval output follows the pattern specified for queries 1 and 2 of Exercise 5-1.

5-4 Derive the equation for the normalized precision measure given in expression (15).

Furnish a construction for the normalized precision similar to the one given in Fig. 5-10 for the normalized recall.

5-5 Assume that the two queries of Exercise 5-1 are retrieved in groups of three items each, that is, the first three items are retrieved together, followed by the next three items, and so on, until the last items are retrieved. Compute the average search lengths obtained for the two queries, assuming the users wish to retrieve 10 documents in all, 15 documents in all, or 20 documents in all.