Evaluation of Coordinate Index Systems During File Development*

DONALD W. KING
Westat Research Analysts, Inc., Bethesda, Maryland
Received November 24, 1964

I. General Comments.—There are three periods in the evolution of an information retrieval system in which evaluation may be helpful. These are (1) the preliminary phase *e.g.*, user studies, preliminary system analysis, development of term lists, etc.); (2) the developmental phase during which time the file is indexed; and (3) the operational phase. This paper is addressed to the problems of evaluation of coordinate index systems during the developmental phase. An indexing experiment, a preliminary search experiment, and an indexing quality control program are presented. Final decisions required during file development may depend, in part, on the results of this experimental work.

II. Evaluation Criteria.—The relative merits of alternative systems and subsystems are expressible in terms of cost, reliability, and time. The relative importance of these factors varies, depending on the needs of the users and the resources available for the system. Thus, some knowledge of these factors and their interrelationships is critical in selecting the "best" system.

System costs can be subdivided into (1) cost of preliminary investigation, (2) cost of indexing, (3) equipment costs, and (4) operating costs. The cost of indexing an entire file can be estimated in an initial indexing experiment. Equipment and operating costs are a function of such characteristics as the total number of documents in the file, the number of terms indexed per document, the total number of terms in the term list, and the number of documents retrieved per search. These characteristics should also be investigated during the initial part of the developmental phase.

Reliability of the system is measured in part by ability to retrieve documents desired by the user. Failure to retrieve desired documents and retrieval of unwanted documents may be attributed to a number of factors such as indexing errors and searching errors. Since indexing errors exist in most systems one must identify them, measure them, relate them to system reliability, and establish a means of controlling them. The evaluation program, described subsequently, shows how these things can be done.

The time factor may be an important constraint on system design and implementation. In the U. S. Patent Office, for instance, a long delay in file development may reduce the effectiveness of the system considerably,[1] since the importance of some areas of invention experience a relatively short life span. Furthermore, search time and retrieval time may be more important in some systems than in others. In the U. S. Patent Office the amount of time available for searching is influenced to a large extent by the production requirements of the office.

Time, cost, and reliability must all be measured in order to provide information necessary for choices among alternative systems.

III. An Evaluation Program.—The principal objectives of evaluation during file development are (1) to select a near optimum indexing procedure, (2) to predict the performance of the system once it is made operational, (3) to provide a means of modifying and correcting the system, if necessary, before the changes become too costly, (4) to control the indexing procedure to ensure that the system will perform satisfactorily when completed, and (5) to serve as a guide in using the system optimally.

The first phase of evaluation during file development involves an indexing experiment (see Figure 1). The indexing experiment is begun after the document file has been essentially completed, the indexers are trained, and a preliminary systems' analysis is undertaken.

The experiment involves indexing a sample of documents by alternative indexing procedures. The documents (or other primary records) are chosen randomly from the entire file. The number of documents chosen for the sample is determined by the usual techniques of experimental design where the answer is dependent upon the desired estimation reliability, the variance of the estimate, and the cost of experimentation.

The principal characteristics of interest in the experiment are indexing accuracy and cost. Another attribute that might be observed is the consistency coefficient[2] between two indexings which is defined as the average of the number of terms selected by both indexers divided by those selected by either of them.

To speak of indexing accuracy implies the ability to determine what is correct. It is not assumed that this can always be done. However, there are instances, such as the indexing of specific chemical compounds, in which it is possible to determine a correct indexing. There are other cases in which a high degree of agreement can be obtained among skilled indexers on the question of whether a particular term (in the general sense) should have been selected. Let us consider the class of indexing environments in which the existence of such a consensus

* Work accomplished under contract to the U. S. Patent Office; paper presented before the Division of Chemical Literature, 148th National Meeting of the American Chemical Society, Chicago, Ill., Sept. 3, 1964.

(1) D. C. Snow, "Coordinate Descriptor Information Systems for Patent Searching," paper presented at the 147th National Meeting of the American Chemical Society, Philadelphia, Pa., April 8, 1964.

(2) J. Jacoby and V. Slamecka, "Indexer Consistency under Minimal Conditions," Documentation Inc., Bethesda, Md., Nov. 1962.

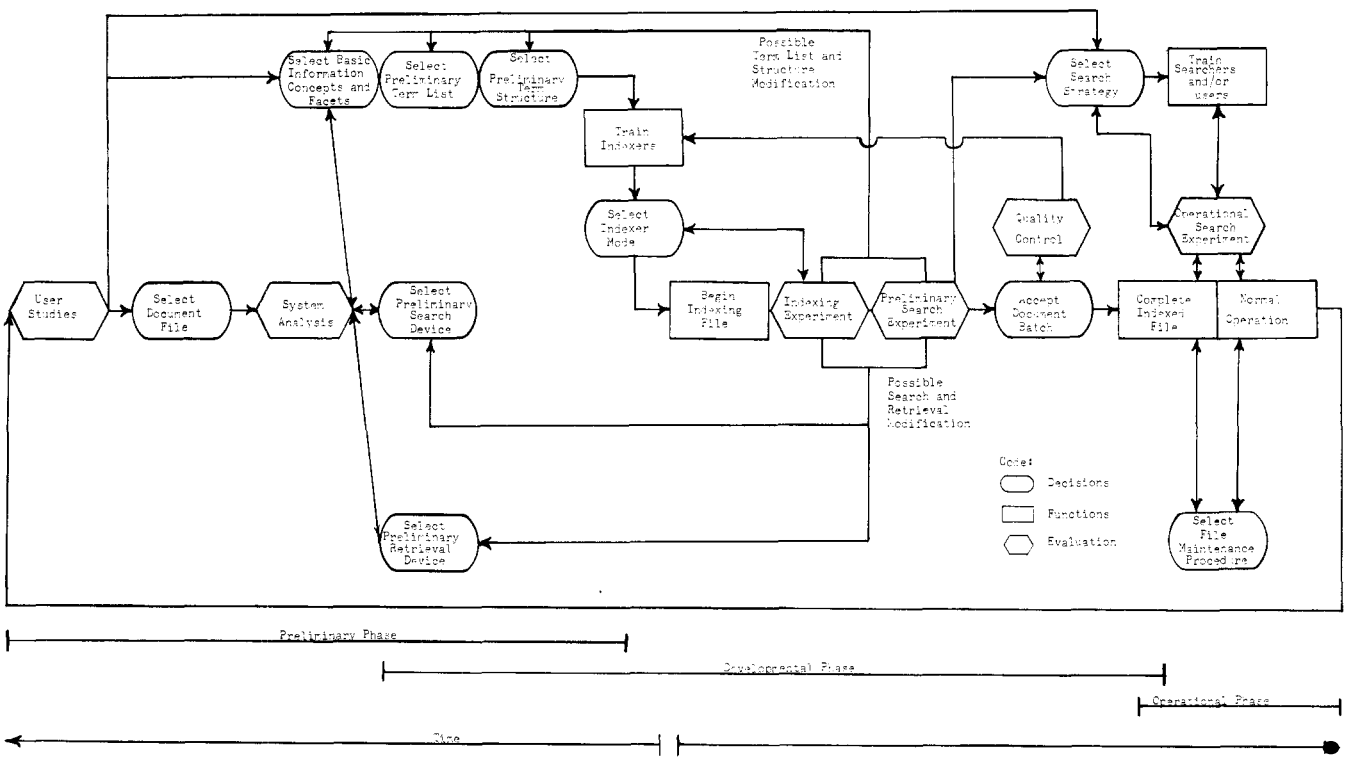

Figure 1.—Flow diagram of decisions, functions, and evaluation of system evolution over time.

can be presumed. We will assume that for a collection of documents and indexing terms such judgments have actually been made, and we will refer to these judgments as the "standard indexing."

The assignment of terms by any indexer may vary from the standard. He may fail to assign terms which should have been assigned and may assign terms which should not have been assigned.

It is clear that if one has a "standard indexing" he can design an experiment which will yield observations on the frequency with which errors in indexing, determined by comparison against the standard, occur. It is also clear that these observations can be aggregated over collections of terms, indexers, and documents, if these aggregations yield meaningful and interpretable results. For such an aggregation, one can define the relative frequencies of Table I.

The $\hat{p}_i$'s are relative frequencies of errors (or correct indexings). If these relative frequencies approach limits as the number of observations becomes large, it is meaningful to interpret these limits as conditional probabilities. For example, if p_2 is such a limit, it is meaningful to interpret it as the probability that a term will be indexed, given that it should not be.

Table I
Categorization of Indexing Errors

	Standard indexing	
Actual indexing	Should not be indexed	Should be indexed
Not indexed	$\hat{p}_0$	$\hat{p}_1$
Indexed	$\hat{p}_2$	$\hat{p}_3$
	$\hat{p}_0 + \hat{p}_2 = 1$	$\hat{p}_1 + \hat{p}_3 = 1$

The results from an experiment[3] with the Organometallics File of the U. S. Patent Office are given in Table II. One can see that this information alone may not be sufficient to decide which indexing procedure is optimum or, indeed, may not be sufficient to decide if any or all procedures are acceptable. To make judgments of this kind, one needs to express certain file output characteristics in terms of indexing errors.

Consider classifying the searched documents into the cell entries of Table III. The cell entries comprise a set of mutually exclusive and exhaustive categories. This retrieval profile is well known. Swets[4] has summarized a number of evaluation criteria involving the observed cell entries of the retrieval profile. The cell entries may be

Table II
Relative Frequencies $\hat{p}_3$ and $\hat{p}_2$ and Time Required for Three Indexing Procedures in the Organometallics File of the U. S. Patent Office[a]

Indexing procedure	$\hat{p}_3$	$\hat{p}_2$	Av. total time per document, min.
Single indexer	0.89	0.0014	64.3
Single indexer reviewed	0.95	0.0002	111.6
Double indexer[b]	0.98	0.0033	128.6

[a] Sample size of 24 documents. [b] Two independent indexings lumped together (set sum).

(3) D. W. King, "Designs of Experiments in Information Retrieval," Proceedings of Social Statistics Section, American Statistical Association Meeting, Cleveland, Ohio, 1963, pp. 103-118.

(4) J. A. Swets, *Science*, 141, 245 (1963).

Table III
Classification of File Documents as a Result of a Search or Searches

Document category	Document Category: Relevant	Not relevant	Total
Retrieved	x_{11}	x_{12}	$x_{1.}$
Not retrieved	x_{21}	x_{22}	$x_{2.}$
Total	$x_{.1}$	$x_{.2}$	$x_{..}$

the number of documents (or information concepts) for a single search (x_{ij}), the average of a number of searches ($\bar{x}_{ij}$), or the distribution of a number of searches ($f(x_{ij})$), depending on the decision requirements.

One would like to relate these cell entries to indexing errors. This can be done in some instances. Consider a k-term search query in which retrieval documents must be indexed by all k terms. Let p_2 and p_3 be "true values" of $\hat{p}_2$ and $\hat{p}_3$ (see Table I) averaged over all indexers, terms, and documents in the file. Assume enough uniformity to make these definitions meaningful. In practice, this appears not to be a critical assumption. Let n_2 and n_3 be the number of terms of the k terms in the search query to which the values p_2 and p_3 apply; $n_2 + n_3 = k$. Let Q_j denote the proportion of the file which *should* contain j of the k terms in the search query. Assume independence of indexing errors from term to term. Then, expected values of the entries in Table III may be obtained

$$\tilde{\bar{x}}_{1.} = x_{..} \sum_{\substack{\text{all values of } n_2, n_3 \\ \text{such that } n_2 + n_3 = k}} p_2^{n_2} p_3^{n_3} Q_{n_3} \qquad (1)$$

$$\tilde{\bar{x}}_{21} = x_{..} Q_k[1 - p_3^k] \qquad (2)$$

$$\tilde{\bar{x}}_{12} = x_{..} \sum_{n_3 < k} p_2^{n_2} p_3^{n_3} Q_{n_3} \qquad (3)$$

from eq. 1–3. If these equations do, in fact, portray the search output, they may be regarded as a model of the retrieval profile. One can obtain estimates of the cell entries by inserting estimates of p_2, p_3, and Q_j in eq. 1 through 3. Other cell entries can be obtained arithmetically. Estimates of these parameters can be obtained from the indexing experiment and the search experiment contemplated in Figure 1.

The preliminary search experiment (see Figure 1) involves indexing a set of sampled documents, in addition to those used in the indexing experiment, and searching this set by a number of search queries. This set of documents can be incorporated into the file with little or no additional costs. The search queries should be as representative as possible of operational search queries. An experiment involving synthetic search queries is discussed in a U. S. Patent Office paper.[5]

(5) E. C. Bryant, D. W. King, and P. J. Terragno, "Analysis of an Indexing and Retrieval Experiment for the Organometallics File of the U. S. Patent Office," WRA PO 10, U. S. Department of Commerce, Patent Office, Aug. 1963.

Estimates of Q_j in eq. 1, 2, and 3 are found by applying the search queries to the standard indexing in the indexing experiment. The validity of the models can be tested by comparing the model estimates of the cell entries with sampling estimates. Comparisons are given for the Organometallics File of the U. S. Patent Office, where the parameters of the model were estimated from an indexing experiment with 24 documents and a search experiment with 201 documents. Both sets of documents were indexed in the same manner. The retrieval profile cell entries are given for queries involving 4, 8, and 12 terms in Table IV. It is clear that this class of models provided good estimates for the retrieval profile in this instance.

Table IV
Comparison of Model Estimates and Observed Cell Entries for the Retrieval Profile of the Organometallics File of 3,625 Documents Prepared by a Double Indexer Procedure

Cell entry	Number of terms per search: 4 terms, Model estimates	4 terms, Sample	8 terms, Model estimates	8 terms, Sample
Total retrieval ($\bar{x}_{1.}$)	48.9	52.7	30.1	32.5
False retrieval ($\bar{x}_{12}$)	3.6	5.1	1.4	2.9
Correct retrieval ($\bar{x}_{11}$)	45.3	47.6	28.7	29.6
Missed documents ($\bar{x}_{21}$)	12.7	9.4	18.4	16.6
Total relevant ($\bar{x}_{.1}$)	58.0	57.0	47.1	46.2

Cell entry	12 terms, Model estimates	12 terms, Sample
Total retrieval ($\bar{x}_{1.}$)	3.6	2.9
False retrieval ($\bar{x}_{12}$)	0	0.7
Correct retrieval ($\bar{x}_{11}$)	3.6	2.2
Missed documents ($\bar{x}_{21}$)	3.8	5.0
Total relevant ($\bar{x}_{.1}$)	7.4	7.2

It should be noted that documents were judged "relevant" if their indexing in the standard set met the search specifications. Thus, the retrieval profile reflects only the effect of indexing errors.

Equations 1 through 3 assume independence of indexing errors from term to term. In case this assumption is unwarranted, one can modify the equations to account for dependencies. Evidently, the assumption of independence was not unwarranted in the illustration cited above.

The decision concerning the optimum indexing procedure is made in view of the "trade off parameters" summarized for three indexing procedures in Table V. Obviously, this information is more valuable than that found in Table II. One can select the optimum indexing procedure with a given number of documents to be indexed, a given number of searches to be conducted per year, and some rough indication of the relative importance of cell entries in the retrieval profile.

The indexing and preliminary search experiments should yield information from which to establish an

Table V
Comparison of Time with Observed Cell Entries for the Retrieval Profile of 184 Search Queries in the Organometallics File of 3,625 Documents

	Indexing Procedure		
	Single indexer[a]	Double indexer[a]	Single indexer reviewed[a]
Indexing time (min.)	64.3	128.6	111.6
Total retrieval ($\bar{x}_{1.}$)	22.7	34.5	26.7
False retrieval ($\bar{x}_{12}$)	4.5	9.2	5.6
Correct retrieval ($\bar{x}_{11}$)	18.2	25.3	21.1
Missed documents ($\bar{x}_{21}$)	9.6	2.5	6.7
Total relevant ($\bar{x}_{.1}$)	27.8	27.8	27.8

[a] These cell entries are averaged over all queries. The cell entries in Table IV are for queries involving 4, 8, and 12 terms only.

acceptable level of indexing accuracy. The indexing quality control program (see Figure 1) is established to ensure that an acceptable level of indexing accuracy is maintained. The QC procedure involves testing indexing accuracy of periodic samples. Quality control acceptance tables can be constructed for various acceptance levels of p_3 and sample sizes. One may wish to test batches of documents prepared over periods of time or by a particular group. Furthermore, one may use these tests to establish a learning curve for future reference.

A more complete discussion of quality control techniques applied to indexing problems may be found elsewhere.[6] Quality control of indexing is particularly important in coordinate index systems since the retrieval errors are highly sensitive to indexing errors.

One must keep in mind that the evaluation procedures described provide specific pieces of information which should be added to a fund of knowledge concerning the entire information retrieval system. Modifications of the search system or retrieval system may be suggested as a result of information found in the evaluations performed during the file development. Need for further training of the indexers may be evident as a result of the indexing experiment. Perhaps the rules of indexing may need modification before indexing is continued. It is even possible that the original system may be abandoned and another approach formulated. The searchers may be instructed to broaden their search queries to reduce retrieval errors attributable to indexing errors.

Thus, the decisions which are made as a result of evaluation during file development can touch every aspect of system operation. It is important that the information upon which these decisions can be based be made available as quickly as possible.

(6) D. W. King, and J. M. Daley, *ADI Proceedings*, 1, 389 (1964).

A Central Information Retrieval System*

D. L. ARMSTRONG and M. T. GRENIER
Aerojet-General Corporation, Azusa, California
Received December 4, 1964

INTRODUCTION

This paper describes a centralized technical information retrieval system for a highly technically oriented company having over 30,000 employees of whom 6,000 are engineers. At the time of the establishment of the system, seven locations were involved. Six were separated by distances ranging from 5 to 400 miles, while the seventh was on the opposite coast.

Simply stated, the problem confronting us was the design and the establishment of an efficient, economical information retrieval system which would permit the maximum utilization of technical information generated at each of the several company plants. It is the purpose of this paper to present our chosen solution to this problem and, in so doing, to emphasize the reason for each decision made.

The several plants of the company are widely separated, and the subject matter includes a broad scope of science and technology. Although each of the several locations has a specific area of responsibility, there are many areas of mutual interest. Exchange of information among these locations did not always occur easily and in a systematic manner even within one plant. Finally, there was no single repository for all technical information. Consequently, the Corporate Technical Information Center (CTIC) was established with the mission of providing a comprehensive index to all company-generated technical information regardless of point of origin and ensuring that new information was brought to the attention of appropriate personnel.

* Presented before the Division of Chemical Literature, 145th National Meeting of the American Chemical Society, New York, N. Y.