A Case Study in the Application of Cranfield System Evaluation Techniques*

SAUL HERNER, F. W. LANCASTER, and WALTER F. JOHANNINGSMEIER
Herner and Company, Washington, D. C.

Received September 18, 1964

This paper is an account of a project performed on behalf of the U. S. Navy Bureau of Ships Technical Library to evaluate and maximize the effectiveness of a computerized information retrieval system based on a specialized thesaurus used in conjunction with the Engineers Joint Council (EJC) system of role indicators and links.[1] The use of links was somewhat limited because of the narrow scope of the documents being indexed, and because of the selectiveness of the indexing procedure. The typical document was indexed under 10 to 15 terms. With role combinations, the number of terms rose to somewhat over 20 per document.

The evaluation method used was that developed by Cleverdon in the ASLIB Cranfield Project.[2] The project began with the indexing of a collection of 750 classified and unclassified reports in the field of Underseas Warfare. In advance of the indexing of the 750 specimen documents, the two indexers assigned to the project took the one-week course given by Battelle Memorial Institute in the EJC system. The purpose of having the indexers take the course was to ensure the greatest possible fidelity to the precepts and purposes of the EJC system, and to eliminate misuse of the system as a factor or variable in the evaluative tests that were to follow.

The indexers used the Bureau of Ships "Thesaurus of Descriptive Terms and Code Book" as their subject authority. This Thesaurus, while considerably more specialized, is similar in content and structure to the "Thesaurus of ASTIA Descriptors." The indexing was selective in that only concepts dealt with centrally and importantly in a report were indexed. The indexing was specific in that the descriptors used to describe concepts were always the most specific available.

The indexing was followed by the translation of the descriptors into alphanumeric code, to facilitate computer manipulation. The indexing, coding, and subsequent storage in the computer, which was an IBM 7090, were followed by a series of test searches based on the collection of 750 specimen documents. The products of the test searches were subjected to exhaustive analyses to measure the retrieval effectiveness of the system and to determine the reasons for failures and less than optimal performances.

Retrieval effectiveness was expressed in terms of relevance and recall ratios. Relevance ratio is defined as the proportion of documents in a search product that are directly responsive to a search topic. Recall ratio is defined as the proportion of known relevant documents in a collection that are actually retrieved in a search. Thus, 100% relevance occurs when no unwanted documents are present in the search product, and 100% recall occurs when all of the documents that are relevant to a question occur in the search product.

Reasons for search failures and less than optimal performance were analyzed in terms of indexing faults, searching faults, and system faults. In all cases where retrieved documents were found to be extraneous to the requirements of a search, or where relevant documents were not retrieved, the reasons were analyzed and defined. By approaching the analyses and definitions in terms of the indexing, the searching, and the system, it was possible to segregate human or operating errors from system or design errors.

The Test Procedure.—The actual test was based on the results of 50 searches. Test questions were produced by a group of scientists and engineers from the staff of the Bureau of Ships. Each member of the group was sent several documents from the test collection. The specimen documents were selected on the basis of the background and activities of each of the individuals involved. Each participant was asked to examine one or more of the specimen documents, and to construct a realistic question to which each would provide a satisfactory answer. Each of the documents upon which questions were based was subsequently referred to as a "source document" for that question.

It is perhaps well to emphasize at this point that, while the questions were directly related to the "source documents" upon which they were based, they were not specifically related to the other documents in the collection.

In all, about 150 questions were collected by this procedure. Of these, 50 were selected for subsequent use in test searches. Formulation of the test searches entailed translation of the natural language question into the language of the system through consultation of the "Thesaurus of Descriptive Terms and Code Book" and through the application of appropriate role indicators. The search program for each question was formulated on a highly comprehensive basis in which series of subqueries were used. The subqueries involved *logical products* (descriptor A *and* descriptor B) and *logical sums* (descriptor A *or* descriptor B). Although the system is capable of handling logical negations and logical sums within subqueries, these devices were not used in the searches.

* Research sponsored in part by the U. S. Navy Bureau of Ships under Contract NObs-88417 and the U. S. Air Force Office of Scientific Research under Contract AF 49(638)-1182; paper delivered at the 148th National Meeting of the American Chemical Society, Chicago, Ill., Sept. 3, 1964.

[1] J. C. Costello, "Training Manual and Workbook for Use in Abstracting and Coordinate Indexing Course," Battelle Memorial Institute, Columbus, Ohio, 1963.

[2] F. W. Lancaster and J. Mills, *Am. Doc.*, 15, 4 (1964).

Table I is an example of a typical set of subqueries. It is designed to search for documents on the testing of "hydrotors" used in starting diesel engines.

Table I

Subquery A	HYDRAULIC ACTUATORS (role 8): DIESEL ENGINES (role 9)
B	HYDRAULIC ACTUATORS (8): DIESEL ENGINES (4)
C	ENGINE STARTERS (9): HYDRAULIC ACCUMULATORS (10)
D	DIESEL ENGINES (4): ENGINE STARTERS (9)
E	ENGINE STARTERS (8): DIESEL ENGINES (4)
F	STARTING (8): DIESEL ENGINES (9): HYDRAULIC ACCUMULATORS (10)
G	ENGINE STARTERS (9): HYDRAULIC SYSTEMS (10)
H	TESTS (8): ENGINE STARTERS (9)
I	DIESEL ENGINES (9): HYDRAULIC ACTUATORS (10)

In essence, this search program involves all major combinations of appropriate descriptors and role indicators. Ordinarily, such blanket searches tend to favor recall and to diminish relevance. However, there is compensation in the fact that the test collection in this case was small and in a relatively homogeneous field. This tends to diminish the likelihood of extraneous documents in the search product.

Analysis for Relevance and Recall.—Once the searches were performed, the documents produced in each case were submitted to the compilers of the original questions. Each compiler was asked to decide for each document whether or not it was reponsive to his question. This furnished a basis for the computation of relevance ratios, or the amount of "noise" or spurious documents in the search product. In cases where spurious or nongermane documents were found, analyses were made to determine the reasons why.

Recall ratios—the proportion of relevant documents in the collection that were present in the search product—were determined by actually doing a total check of the specimen collection to locate all documents that had any possible relevance to ten of the test questions. All seemingly relevant documents that were not uncovered in the original searches were submitted to the question compilers for relevance assessment. For any of these additional documents judged to be relevant, further analyses were made to find out why they had been missed in the original computer search.

Test Results.—Table II summarizes the results of the computations of the relevance and recall ratios. These figures, when compared with the results of similar tests on other systems, indicate a situation in which a large percentage of the most germane documents are being missed, but in which those documents that *are* being produced in a search have a relatively high likelihood of being directly responsive to the search topic. Typical relevance and recall ratios at Cranfield were in the vicinity of 20% relevance and 80% recall.[2]

Table II

	Relevance, %	Recall, %
Source + relevance A documents (as relevant as source documents)	29.7	68.3
Source + relevance A documents + relevance B documents (less relevant than source documents)	54.3	53.8

Table III summarizes the main reasons for the retrieval of nonrelevant documents or for the production of "noise" or spurious documents in the searches that were done.

From Table III, it is clear that by far the greatest cause of "noise" in the search results was searching errors, and the most frequently occurring type of searching error was the failure to select crucial terms or concepts in translating the questions into the language of the system.

Table III

Cause	No. of unwanted items retrieved
Searching	
Searches too generic	60
Searches too specific	4
Omitted important concept demanded in question (*i.e.*, not enough qualifying descriptors used)	122
Total searching failures	186
Indexing	
Indexer used inappropriate descriptor	1
Total indexing failures	1
System	
Vocabulary not sufficiently specific	11
Role indicator problems	
Indexer chose inappropriate role	1
Total system failures	12
Total nonrelevant documents retrieved	199

The story was somewhat different when we analyzed for reasons for failure to retrieve documents that were relevant to search questions (Table IV). Here, indexing errors contributed more heavily to the cause of failure. System inadequacies were also more significant contributing factors. Errors in searching procedures were still important

Table IV

Cause	Number of documents missed
Indexing	
Omitted important concept	12
Not sufficiently specific	1
Total indexing failures	13
Searching	
Omitted important concept	5
Subsearchers demanded co-occurrence of too many descriptors	2
Searcher too specific	1
Searcher did not exhaust reasonable search possibilities	12
Total searching failures	20
System	
Role indicator problems	
Searcher did not exhaust possible roles	4
Indexer did not use all appropriate roles	5
Insufficient linkages in thesaurus	2
Error in question coding	2
Total system failures	13
Total failures to retrieve known relevant documents	46

in recall failures, but not so much as in the case of "noise" or spurious documents in the search product.

At this point, a parenthetical explanation is perhaps in order to account for the large number of nonrelevant documents produced and the seemingly small number of relevant documents missed in a system that is supposed to be high on relevance and low in recall. The explanation lies in the fact that the figures for relevance are based on all 50 searches, whereas the figures for recall are based on only 10 searches.

Tables V and VI are analyses of two of the 50 test searches that were done, with explanation of failures where they occurred. These two analyses were chosen because they are extreme examples, and they furnish us an opportunity to illustrate indexing, searching, and system breakdowns.

Table V
Question 41. Underway Evaluation of Vibratory Characteristics for Surface Ship Appendages

Documents retrieved	Documents missed
7 relevant documents 20 nonrelevant documents	2 relevant documents
Relevance ratios	Recall ratios
7/27 = 26%	7/9 = 78%

Reasons for Nonrecall of Relevant Documents

Role indicator problems. One document appears to have been missed for a joint reason: the indexer did not use all appropriate roles, and the searcher appears to have chosen an inappropriate role. The searcher asked for VIBRATION (9): PROPELLERS (MARINE) (4) whereas the indexer had used the form VIBRATION (7): PROPELLERS (MARINE) (9). The searcher's use of ROLE 4 with PROPELLERS (MARINE) in this context does not seem the most likely one to retrieve documents in answer to this question. ROLE 9 (*i.e.*, vibration OF or IN propellers) seems more appropriate. On the other hand, the indexer failed to use ROLE 9 with vibration although harmonic analysis was being applied to study the vibration.

Indexer omitted an appropriate role. The second relevance A document was lost because the searcher used VIBRATION (9): SHIP STRUCTURAL COMPONENTS (9), whereas the item had been indexed as VIBRATION (9): SHIP STRUCTURAL COMPONENTS (4). The indexer appears to have used ROLE 4 in the sense that a mathematical analysis of vibrational characteristics has been derived which is applicable to ship structural components. On the other hand, the indexer did not use ROLE 9 as being appropriate for a component undergoing vibration.

Reasons for Retrieval of Nonrelevant Documents

Searcher too generic. Twenty unwanted items were retrieved by subsearches in which the precise requirements of "surface ships" and "appendages" were neglected. The subsearches asked for SHIP STRUCTURAL COMPONENTS (9) with either VIBRATION (9) or SHIP NOISE (9). The unwanted items either dealt with submarines or contained no information on appendages.

In Question 41, Table V, reasons for failures to retrieve relevant documents revolved around the misapplication of role indicators by both the indexers and the searchers. This was a fairly common problem, and points up the need for better and clearer definitions of the role indicators if they are to serve as useful tools rather than deterrents to effective searching.[3]

The reason for the production of spurious documents in response to Question 41 was that the searcher was not specific enough in her selection of search terms. The

Table VI
Question 44. Information on Optimum Machinery Foundation Design for Attenuating Vibrations from Machinery to Hull Structures

Documents retrieved	Documents missed
1 nonrelevant document	13 relevant documents
Relevance ratios	Recall ratios
Zero relevance	Zero recall

Reasons for Nonrecall of Relevant Documents

Searcher omitted important concept demanded in the question. She completely neglected the concept of "hulls" although it is explicit in the question. If she had asked for HULLS (MARINE) (9): VIBRATION (9, 8, or 7) she would have retrieved the source document, three relevance A documents, and a relevance B item.

Searcher did not exhaust reasonable search possibilities. The questioner asked for "vibrations from machinery," implying ship's machinery. The searcher used only the term MACHINES to cover this concept. This is not a descriptor appropriate to cover the chief types of marine machinery. If she had tried the two most obvious sources of marine machinery vibrations, namely, MARINE ENGINES (9) and REDUCTION GEARS (9), with the term VIBRATION (9), three relevance B documents would have been captured.

Indexer omitted important concept. A relevance B document was missed because the indexer omitted the idea of "reduction," although the concept of "noise reduction" was an important one dealt with in the document.

A relevance A document, missed in searching, deals with the effect of shock on foundations. A principal effect of the shock loading would be to set up vibrations in the foundation. The indexer missed this concept of "vibration," although in fairness the concept of vibration is mentioned in only a single paragraph.

For another relevance B document, the indexer omitted an appropriate descriptor, DESIGN. However, even if she had included the term DESIGN (8), the document would not have been retrieved, since she had used the term VIBRATION ISOLATORS in roles 8 and 10, whereas the searcher asked for DESIGN (8): VIBRATION ISOLATORS (9).

Role indicator problems. Searcher did not exhaust possible roles. A relevance B document was lost when the searcher asked for MACHINERY NOISE (9): REDUCTION (9), whereas the indexer had used MACHINERY NOISE (9): REDUCTION (4 or 8). The searcher was largely to blame because role 8 seems somewhat more appropriate here than role 9. However, the indexer could have used role 9 also.

Insufficient linkages in thesaurus. One relevance B document was missed because it was indexed precisely as ENGINE NOISE whereas the searcher asked for the term MACHINERY NOISE. Although these two terms are closely related (in fact "engine noise" may be regarded as a species of "machinery noise") they are not linked in the thesaurus. If they had been linked, the searcher may have been led to use the term ENGINE NOISE.

Reasons for Retrieval of Nonrelevant Documents

Searcher omitted important concept demanded in the question (*i.e.*, "foundations").

[3] F. W. Lancaster, *Spec. Libraries*, 55, 696 (1964).

question dealt with vibratory characteristics of surface ship appendages, but the searcher asked for Ship Structural Components combined with either Vibration or Ship Noise. She neglected to specify Surface Ships and Appendages. As a result, she got out a large number of documents on submarines, together with a group that dealt with surface ships but not their appendages. This points up a need for a clearer indication of relationships among specific terms and between generic and specific terms in the Thesaurus. It also points up a need for better coordination between indexing and searching policy and procedures.

The second question, Question 44, Table VI, was a total failure. The reasons for failure were many and varied. First of all, regarding failure to recall relevant documents, the contributing factors were the following: the person setting up the search omitted a basic concept from her search program, the searcher failed to exhaust all possible subjects and subject combinations, the indexer omitted an important concept, the indexer did not apply all of the appropriate roles to the descriptors under which the documents should have been retrieved, and the Thesaurus failed to establish relationships among vital terms and caused them to be missed on both the searching and indexing sides.

Regarding the retrieval of nonrelevant documents, the prime cause was that the searcher failed to use an important qualifying subject in her search program. Thus, we see in Question 44 examples of indexing, searching, and system failures. While, for reasons of time and space, they are not expounded in detail in this paper, the causes underlying each specific failure and class of failure were analyzed in detail for each test question, and remedial steps were put forth wherever possible.

Implications and Conclusions.—Having reviewed the essentials of the test procedure, we can now turn to the question of its significance. What can one reasonably expect to derive from this type of test? Perhaps a good place to start answering this question is by stating what one *cannot* expect to derive from this type of test. Relevance and recall ratios, as we have used them, cannot be construed as figures of merit; they do not tell us whether we have a good or bad system in any absolute sense. What they do tell us is what kind of system we have, and it is for us to decide whether what we have meets our needs.

In the case of the Bureau of Ships system, we have a situation of high relevance and low recall, meaning that the average search is likely to produce a relatively small core of highly relevant documents but is also likely to miss a significant number of relevant documents. Depending on one's view and requirements, this highly or overly purified type of search product could be very desirable. Libraries in the past have been criticized for being overly expansive and not discriminating enough in their searches, with the result that they have answered requests with large amounts of material, much of which is of questionable relevance to the needs of the requestor.

The Bureau of Ships system clearly gets around this problem by furnishing a highly purified search product. However, it does so at the expense of completeness. If it is determined that completeness is the desired goal, then the system has to be changed. As we mentioned earlier, the indexing in the Bureau of Ships project was both specific and selective. In choosing terms from the Thesaurus, the indexers went after the most specific manifestation of each important concept uncovered in a document. The indexers were also highly discriminating in the concepts they indexed, choosing only those that were dealt with centrally and importantly in a document.

This approach makes for high relevance and low recall. If it were determined that purity is less important than completeness, the situation at the Bureau of Ships could be altered by lessening discrimination and specificity in the selection of concepts and index terms to describe them. Documents would be indexed under all significant concepts on which they touch, regardless of how much information they convey about them; concepts would be indexed by less specific and more generic terms. This would increase recall of relevant documents, but would also produce more extraneous or unwanted documents. This increase in recall and decrease in relevance could also be attained through the use of fewer role indicators in either or both indexing and searching. If increased recall were to be attained through more exhaustive indexing, the noise level could be diminished through a greater use of links, which were not generally warranted with the selective type of indexing that was done in the present case.

Leaving the question of desired completeness of the search product, another important thing that the evaluation technique under discussion can tell us about a system is where it falls down in its design or implementation, and how these shortcomings can be remedied. In the course of this paper, we have discussed search failures due to indexing errors, errors in searching procedures, and basic faults in the system itself. In every case where errors are encountered, they are carefully characterized as to nature and cause. This type of detailed analysis furnishes a basis for remedy and correction.

In all such cases of change or correction, regardless of whether the changes involve deeper or less specific indexing to produce greater recall, or whether they involve alterations in indexing techniques or indexing authorities to produce greater accuracy, relevance and recall ratios can be extremely useful. They serve as indices against which the effects of system or operational changes can be measured. A single set of relevance and recall ratios, backed by detailed analyses of causative factors, can give us a general picture of what a system is doing and some indications of why it is doing it. But when we compare a single set of ratios with a second set based on altered system components or operating procedures, we have a powerful tool for determining whether and to what degree we are going in the direction of the system we want. However, no evaluation technique can tell us what we want or need. This we have to decide for ourselves.