PERFORMANCE OF RUECKING'S WORD-COMPRESSION METHOD WHEN APPLIED TO MACHINE RETRIEVAL FROM A LIBRARY CATALOG

Ben-Ami LIPETZ, Peter STANGL, and Kathryn F. TAYLOR:
Research Department, Yale University Library, New Haven, Connecticut

F. H. Ruecking's word-compression algorithm for retrieval of bibliographic data from computer stores was tested for performance in matching user-supplied, unedited bibliographic data to the bibliographic data contained in a library catalog. The algorithm was tested by manual simulation, using data derived from 126 case studies of successful manual searches of the card catalog at Sterling Memorial Library, Yale University. The algorithm achieved 70% recall in comparison to conventional searching. Its acceptability as a substitute for conventional catalog searching methods is questioned unless recall performance can be improved, either by use of the algorithm alone or in combination with other algorithms.

Frederick H. Ruecking has published a report (1) of a method for improving bibliographic retrieval from computerized files when searching on unverified input data supplied by requestors. The method involves compression of author-and-title information before comparison. The rules for compression cause certain types of spelling errors and word discrepancies to be ignored by the computer. Ruecking reported 90.4% recall and 98.67% accuracy (precision) in a test of his method in which unverified book order requests were matched against a MARC I data base that contained 1392 of the references searched. This paper reports on a small-scale manual simulation test undertaken to assess the value of the method when applied to bibliographic retrieval from a library catalog.

The opportunity to test Ruecking's method when applied to retrieval from a library catalog was provided by the ready availability of data derived from a current study (2) of catalog use at Sterling Memorial Library (3.5 million books) at Yale University. This study collects, from a rigidly randomized sample of catalog users, precise information on the clues available to them at the moment of initiating a search. Search clues are recorded exactly as known to the catalog user, employing his own spelling—right or wrong. For each catalog user studied, the outcome of the search is ascertained; complete catalog information is recorded for documents identified as pertinent in successful searches. Search clues known to catalog users who seek specific documents correspond to the "unverified input data" which Ruecking's method would match against catalog holdings. Catalog information on those documents identified as pertinent corresponds to the portion of the data base that Ruecking's program seeks to match. It was possible, therefore, to apply Ruecking's method by manual simulation, and to test its recall performance in real catalog searches. A test of its precision was not immediately feasible because such a test would require comparison of input data with the entire catalog (or a substantial portion of it). However, the determination of recall performance would at least indicate whether the method shows sufficient promise in catalog searching to warrant evaluation of its precision.

An aside on precision is in order, however. It should be noted that precision of retrieval with a given method tends to vary inversely with the size of the file being searched. Although Ruecking did not specify the number of records included in his MARC I data base, it could not have exceeded 48,000. Had he run his test on a data base, ten, or fifty, or one hundred times larger, the measured precision would certainly have been much lower than the figure reported. Any librarian who is contemplating the adoption of a retrieval technique which has been tested on a data base similar to, but smaller than, his own should realize that precision performance must inevitably drop as the data base is increased. The degree of lowered precision to be expected may be predicted theoretically or estimated from tests on files of several different sizes.

The data used in the evaluation of recall performance reported in this paper came from 126 searches in which the catalog users had been successful in locating the specific documents that they were seeking. The compression coding method described by Ruecking was applied in each instance to the author-title search clues supplied by the catalog user and to the author-title information available on the catalog card. Threshold values were computed for the catalog card data, and retrieval values were computed for the user data. When the retrieval value was at least as large as the threshold value, the document was considered "retrieved."

Ruecking's method was designed for use with English-language titles only. Of the 126 catalog searches in the study sample, 20 involved foreign-

language titles. Recall was determined on both the full sample and the English-language subset of 106 searches. Surprisingly, there is not a great improvement in performance when foreign-language references are excluded.

It should be noted that several difficulties were encountered in applying Ruecking's method because of ambiguities in the rules stated in his paper. In fact, in his Figure 2 (page 236), of the seventeen illustrations of compression-coded data retrieved by his program, at least eight appear to contain departures from the compression-coding rules as stated in the paper. His Table 5 (page 235) is scantily described: "Individual Code Test" and "Full-Code Test" are not defined; neither are column headings. And, contrary to the text (page 234), values in columns five through seven are obtained by adding two to the calculated thresholds in only the top half of Table 5; in the bottom half, no such regular correlation exists. In all cases of ambiguity, the alternative was selected that would tend to increase probability of retrieval. For example, Ruecking states (page 234) that the search program provided for matching of titles on the basis of rearrangement of title words, and that the threshold value required for retrieval is raised at the same time. Raising this value decreases the probability of retrieval, but it is not clear by how much the value is to be raised. For purposes of the test, the threshold value was not raised at all in cases where title words were out of correct sequence, thus retaining maximum probability of retrieval based on the number of matched words alone, regardless of their sequence.

Results of the test showed that, of the 126 documents in the full sample which were located successfully by manual search in the existing card catalog, only 88 were retrieved by the compression-code method—a recall rate of 70%. Considering only the 106 English-language references, 77 were retrieved by the compression-code method—a recall rate of 73%.

The premise for the preceding calculation of recall rate should be clearly understood. The test considered real document searches that were concluded successfully in an actual library using a manual catalog; recall is defined here as the proportion of such searches that would be concluded successfully in a hypothetical, computerized library where the only means of searching the catalog would be by Ruecking's method. In a real library with a manual catalog, wanted documents can be located in many ways, not merely through a knowledge of author and title (e.g., through subject entries, series entries, cross references). The test did not disqualify any manual approaches from consideration; it compared the real world with a specific potential alternative. Obviously, the use of Ruecking's method in combination with other computer programs could result in a recall rate higher than 70% or 73% by the method of calculation employed, and conceivably higher than 100% (because some document searches of manual catalogs that now end in failure might become successful using new search methods).

Table 1 provides detailed information on the discrepancies between user data and catalog data in the test. With respect to the full sample (126 documents), there were 49 documents for which mismatches of data were observed. Of these, the compression-code method was able to "heal" mismatches in 11 instances to cause retrieval; on the other hand, manual searches had achieved retrieval in all 49 instances. With respect to the English-language sample (106 documents), there were 37 documents for which mismatches of data were observed. Of these, the compression-code method was able to "heal" mismatches in 8 instances to cause retrieval; on the other hand, manual searches had achieved retrieval in all 37 instances.

Contrary to expectations, the compression-code method performed somewhat worse, or at least no better, in "healing" actual mismatches in English references (8 out of 37) than it did with foreign-language references (3 out of 12). The higher overall recall percentage with the English-

Table 1. Results of Applying Ruecking's Method in Cases where User Clues and Catalog Data Did not Match Completely

	Full Sample (126 documents)		*English Subset (106 documents)*	
Type of Mismatch in User Data	*Retrieved*	*Not Retrieved*	*Retrieved*	*Not Retrieved*
Had neither author nor title		2		1
Had author's last name, no title		9		5
Had title, no author	1	2	1	2
Had wrong author		1		1
Had misspelled author	4	4	2	1
Had wrong words in title	1	9*	1	6
Had misspelled words in title	2	2	1	2
Had words transposed in title	2		2	
Had incomplete title:				
a. First word correct	2	5**	2	5**
b. First word incorrect		6		5
Had entire subtitle, no title		1		1
Had part of subtitle				
a. First word correct		1		1
b. First word incorrect		2		2
Total documents***	11	38	8	29

*1 case of correct word stems not matched because of wrong endings.

**2 cases of long or composite titles with maximum threshold values contained in input words but not among the first four significant words.

*** Figures shown are lower than totals of figures in columns because some documents had two or more types of mismatch.

language subset is attributable entirely to the fact that users had complete and correct data more frequently for English references (69 out of 106) than they did for foreign-language references (8 out of 20). Thus, regardless of original intent, the method words equally well (or equally poorly, depending on one's viewpoint) on foreign-language and English references. If foreign-language references had been systematically ignored in applying the test to catalog searches, some 16% (20 out of 126) of the searches would have been excluded, with no real gain in performance.

The block of interviews from which the searches used in this test were drawn included 10 unsuccessful document searches in addition to the 126 successful searches. One could speculate on whether the compression-code method would have been able to "heal" these failures, resulting in a higher performance rating. The indications are, however, that the chances of such healing are close to zero. In a majority of these unsuccessful searches, the available data were incomplete or were not of the type that the method is intended to utilize. In the few remaining cases, it is very likely that the searches were unsuccessful simply because the desired documents were not in the library collection.

Recall performance as measured by the test could have been improved by modifying Ruecking's rules to some extent. For example, five more titles would have been retrieved had the assigned retrieval value been increased by two units in cases where the first title word matched correctly; this would have increased overall recall performance from 70% to 74%. A further increase to 76% would have resulted from matching the user's version of the title with the catalog's subtitle, or with portions of titles which follow a punctuation mark (in addition to matching with the actual title in the catalog).

Extension of the compression code to include publisher and date as well as author and title would do little or nothing to improve the performance of this method. The test data, although admittedly a small sample, indicate that users who do not have accurate author and title information when they begin a search very rarely have accurate information on any other descriptive data element.

It is, of course, a matter for individual judgment as to whether the performance of the compression-code method, as indicated by the test reported here, is sufficiently good to make it attractive for use in some computerized alternative to the manual library catalog. In the authors' opinion, Ruecking's method does not in itself supply an adequate solution to the problem of searching a computerized catalog. However, further investigation seems warranted along two lines. First, the method might be modified to give better performance in this application. Second, it might be used in combination with some other computer methods to give searching performance approaching that which is attained today by the manual searching of card catalogs.

ACKNOWLEDGMENT

The work reported in this paper was supported in part by a grant from the U.S. Office of Education, OEG-7-071140-4427.

REFERENCES

1. Ruecking, Frederick H., Jr.: "Bibliographic Retrieval from Bibliographic Input; The Hypothesis and Construction of a Test," *Journal of Library Automation, 1* (December 1968), 227-38.
2. Lipetz, Ben-Ami; Stangl, Peter: "User Clues in Initiating Searches in a Large Library Catalog," in *American Society for Information Science, Proceedings, 5.* Annual Meeting, October 20-24, 1968, Columbus, Ohio, p. 137-139.

BOOK REVIEWS

Conceptual Design of an Automated National Library System, by Norman R. Meise. Metuchen, N.J.: Scarecrow Press, 1969. 234 pp. $5.00.

This is a very confusing book. And it is too bad, because this reviewer kept feeling that the author, Norman Meise, had something to present. The trouble is that he does not communicate. This, I think, is the result of two things. First, the book reflects the naiveté of engineers when they come to deal with what are basically social systems like libraries. This does not mean it can't be done, but such a task needs clarity and purpose, which this book does not have. The second springs from this failure. The masses of data, assumptions, and commentary in the book are poorly organized and interrelated. It is not enough to write strings of words; those strings must communicate and relate backward and forward in the text.

Although never explicitly stated, the book evidently grew out of a study performed by the United Aircraft Corporate Systems Center in 1965-66 for the development and implementation of a Connecticut Library Research Center (see ERIC Document ED 0221512). The latest reference in the book is 1966. In a field, i.e. library networks, where a fair amount of work and discussion has taken place in the last three years (e.g. the EDUNET Conference in 1966), a book like this quickly loses its impact.

The purpose of the book, according to the author, is "to show the feasibility of a system concept rather than provide a detailed engineering design." The system is "an automated national library system" using the State of Connecticut as a model. The author then adds (spoiling the whole introduction): "If these functions (bibliographic searching, acquisition, cataloging, circulation) can be economically automated, the major problems associated with our information explosion will be solved." As Anatole France once said: "It is in the ability to deceive oneself that the greatest talent is shown."