TITLE-ONLY ENTRIES RETRIEVED BY USE OF TRUNCATED SEARCH KEYS

Frederick G. KILGOUR, Philip L. LONG, Eugene B. LIEDERMAN, and Alan L. LANDGRAF: The Ohio College Library Center, Columbus, Ohio.

An experiment testing utility of truncated search keys as inquiry terms in an on-line system was performed on a file of 16,792 title-only bibliographic entries. Use of a 3,3 key yields eight or fewer entries 99.0% of the time.

A previous paper (1) established that truncated derived search keys are efficient in retrieval of entries from a name-title catalog. This paper reports a similar investigation into the retrieval efficiency of truncated keys for extracting entries from an on-line, title-only catalog; it is assumed that entries retrieved would be displayed on an interactive terminal.

Earlier work by Ruecking (2), Nugent (3), Kilgour (4), Dolby (5), Coe (6), and Newman and Buchinski (7) were investigations of search keys designed to retrieve bibliographic entries from magnetic tape files. The earlier paper in this series and the present paper investigate retrieval from on-line files in an interactive environment. Similarly, the work of Rothrock (8) inquired into the efficacy of derived truncated search keys for retrieving telephone directory entries from an on-line file.

Since the appearance of the previous paper, the Ohio State University Libraries have developed and activated a remote catalog access and circulation control system employing a truncated derived search key similar to those described in the earlier paper. However, OSU adopted a 4,5 key consisting of the first four characters of the main entry and the first five characters of the title excluding initial articles and a few other nonsignificant words. Whereas the OSU system treats the name and title as a continuous string of characters, the experiments reported in this and the previous paper deal only with the first word in the name and title, articles always being excluded.

The Bell System has also recently activated a Large Traffic Experiment in the San Francisco Bay area. The master file in this system contains 1,300,000 directory entries. The system utilizes truncated derived keys like those investigated in the present experiments.

MATERIALS AND METHODS

The file used in this experiment was described in the earlier paper (1), except that this experiment investigates the title-only entries. The same programs used in the name-title investigation were used in this experiment; the title-only entries were edited so that the first word of the title was placed in the name field and the remaining words in the title field.

As was the case formerly, it was necessary to clean up the file. Single word titles often carried in the second or title field such expressions as ONE YEAR SUBSCRIPTION or VOL 16 1968. In addition there were spurious character strings that were not titles, and in such cases the entire entry was removed from the file. Thereby, the original 17,066 title entries were reduced to 16,792.

The truncated search keys derived from these title-only entries consist of the initial characters of the first word of the title and of the second word of the title. If there was no second word, blanks were employed. If either the first or second word contained fewer characters than the key to be derived, the key was left-justified and padded out with blanks.

To obtain a comparison of the effectiveness of truncated research keys derived from title-only entries as related to first keys derived from name-title entries, a name-title entry file of the same number of entries (16,792) was constructed. A series of random numbers larger than the number of entries in the original name-title file (132,808) was generated and one of the numbers was added to each of the 132,808 name-title entries in sequence. Next the file was sorted by number so that a randomized file was obtained. Then the first 16,792 name-title entries were selected. The same program analyzed keys derived from this file.

RESULTS

Table 1 presents the maximum number of entries to be expected in 99% of replies for the file of 16,792 title-only entries as well as for the name-title file containing the same total of entries. For example, when a large number of random requests are put to the title-only file using a 3,3 search key, the prediction is that 99.0% of the time, eight or fewer replies will be returned. However, in the case of the name-title file, only two replies will be returned 99.3% of the time.

The 3,3 key produced only thirteen replies (.12% of the total number of 3,3 keys) containing twenty-one or more entries. The highest number of entries for a single reply for the 3,3 key was 235 ("JOU,OF" derived from JOURNAL OF). The next highest number of replies was 88 ("ADV, IN" for ADVANCES IN).

Table 1. Maximum Number of Entries in 99% of Replies

Search Key	*Title-Only Entries*		*Maximum Entries*	
	Name-Title Entries Per Reply	*Percent Of Time*	*Maximum Entries Per Reply*	*Percent Of Time*
2,2	15	99.1	7	99.0
2,3	12	99.1	4	99.6
2,4	11	99.0	3	99.5
3,2	9	99.1	3	99.2
3,3	8	99.0	2	99.3
3,4	8	99.1	2	99.5
4,2	8	99.1	2	99.2
4,3	7	99.0	2	99.6
4,4	7	99.1	2	99.7

DISCUSSION

The two words from which the keys are derived in name-title entries constitute a two-symbol Markov string of zero order, since the name string and title string are uncorrelated. However, the two words from which keys are derived in the title-only entry are first order Markov strings, since they are consecutive words from the title string and are correlated. The consequence of these two circumstances on the effectiveness of derived keys is clearly presented in Table 1. The keys from name-title entries consistently produce fewer maximum entries per reply. Therefore, it is desirable to derive keys from zero order Markov strings wherever possible.

The Ohio State University Libraries contain over two and a quarter million volumes, but on 9 February 1971 there were only 47,736 title-only main entries in the catalog. The file used in the present experiment is 35% of the size of the OSU file. Since 99% of the time the 3,3 key yields eight or fewer titles, it is clear that such a key will be adequate for retrieval for library on-line, title-only catalogs.

The 3,3 key also possesses the attractive quality of eliminating the majority of human misspelling as pointed out in the earlier paper (1).

There remains, however, the unsolved problem of the efficient retrieval of such titles as those beginning with "Journal of" and "Advances in". It appears that it will be necessary to devise a special algorithm for those relatively few titles that produce excessively high numbers of entries in replies.

In the previous investigation it was found that a 3,3 key yielded five or fewer replies 99.08% of the time from a file of 132,808 name-title entries. Table 1 shows that for a file of only 16,792 entries the 3,3 key produces two or fewer replies 99.3% of the time. These two observations suggest that as a file of bibliographic entries increases, the maximum number of entries per reply does not increase in a one-to-one ratio, since the maximum number of entries rose from two to five while the total size of the file increased from one to approximately eight. Further research must be done in this area to determine the relative behavior of derived truncated keys as their associated file sizes vary.

CONCLUSION

This experiment has produced evidence that a series of truncated search keys derived from a first order Markov word string in a bibliographic description yields a higher number of maximum entries per reply than does a series derived from a zero order Markov string. However, the results indicate that the technique is nonetheless sufficiently efficient for application to large on-line library catalogs. Use of a 3,3 search key yields eight or fewer entries 99.0% of the time from a file of 16,792 title-only entries.

ACKNOWLEDGMENT

This study was supported in part by National Agricultural Library contract 12-03-01-5-70 and by Office of Education contract OEC-0-72-2289 (506).

REFERENCES

1. F. G. Kilgour; P. L. Long; E. B. Leiderman: "Retrieval of Bibliographic Entries from a Name-Title Catalog by Use of Truncated Search Keys," *Proceedings of the American Society for Information Science* 7 (1970), pp. 79-82.
2. F. H. Ruecking, Jr.: "Bibliographic Retrieval from Bibliographic Imput; The Hypothesis and Construction of a Test," *Journal of Library Automation* 1 (December 1968), 227-38.
3. Nugent, W. R.: "Compression Word Coding Techniques for Information Retrieval," *Journal of Library Automation* 1 (December 1968), 250-60.
4. F. G. Kilgour: "Retrieval of Single Entries from a Computerized Library Catalog File," *Proceedings of the American Society for Information Science* 5 (1968), pp. 133-36.
5. J. L. Dolby: "An Algorithm for Variable-Length Proper-Name Compression," *Journal of Library Automation* 3 (December 1970), 257-75.
6. M. J. Coe: "Mechanization of Library Procedures in the Medium-sized Medical Library: X. Uniqueness of Compression Codes for Bibliographic Retrieval," *Bulletin of the Medical Library Association* 58 (October 1970), 587-97.
7. W. L. Newman; E. J. Buchinski: "Entry/Title Compression Code Access to Machine Readable Bibliographic Files," *Journal of Library Automation* 4 (June 1971), 72-85.
8. H. I. Rothrock, Jr.: *Computer-Assisted Directory Search; A Dissertation in Electrical Engineering*. (Philadelphia: University of Pennsylvania, 1968).