Corporate Author Entry Records Retrieved by Use of Derived Truncated Search Keys

Alan L. LANDGRAF, Kunj B. RASTOGI, and Philip L. LONG,
The Ohio College Library Center.

An experiment was conducted to design a corporate author index to a large bibliographic file. The nature of corporate entries necessitates a different search key construction from that of personal names or titles. Derivation of a search key to select distinct corporate entry records is discussed.

INTRODUCTION

This paper describes the findings of an experiment conducted to design a corporate author index to entries in a large file of catalog records at the Ohio College Library Center; a companion paper describes findings of a similar investigation into retrieval employing a personal author index.[1] The center has operated an on-line, shared cataloging system since August 1971. In addition to a Library of Congress card number index, the system maintains truncated name-title and title index files. The user is thus able to retrieve entries employing truncated search keys. Three previous papers report results of experiments which led to the design of the name-title and title indexes.[2-4]

For monographs having personal names as main entries, a truncated 3,3 search key consisting of the first three letters of the author's name plus the first three letters of the first non-English-article word of the title was judged to be satisfactory in that this key yielded five or fewer entries per query in more than 99 percent of the cases when keys were selected at random.[5] However, a recent study by Guthrie and Slifko reveals that a model which employs random selection of entries yields results closer to actual experience, and with a higher average number of entries per reply.[6]

A search key composed of the first five or four characters of the surname and the first or first and second initials makes possible efficient retrieval.[7] However, the situation is different in the case of corporate entries because many corporate names begin with the same or similar words. For example, in the records examined, the initial words of more than 1,300 publications are "U.S. Congress, House Committee On. . . ." Obviously a type of search key different from that which proved efficient for retrieving personal authors is required for retrieval of corporate entries.

MATERIAL AND METHODS

The experiment used a file of approximately 200,000 MARC II records having a total of 68,169 corporate name entries. Corporate entries were extracted from the 110, 111, 410, 411, 710, 711, 810, and 811 fields in the records. A program edited the file to extract keys; initial English language articles were removed from each entry, and the words "United States," "U.S.," "U. S.," "Great Brit.," and "Great Britain" appearing anywhere in the entry were replaced with "US" and "Gt Brit" respectively. A blank was substituted for each subfield delimiter and associated code, and unwanted characters such as punctuation, diacritics, and special symbols were removed; the program also closed up the space that the unwanted character had occupied. One blank replaced multiple blanks. The elements extracted consisted of five segments of eight characters each, representing the initial eight characters of the first five words of the corporate entry. Segments containing fewer than eight characters were padded out with blanks. If a corporate name had fewer than five words, the remaining segments were blank.

To study a given type of key, the file was sorted on a specified number of initial characters of each segment; these initial characters were then employed as search keys by a program which sequentially compared the characters in the key, counting distinct and identical keys.

RESULTS AND DISCUSSION

Table 1 presents the number of distinct keys and the maximum number of occurrences of identical keys for the structures studied in the experiment. The larger the number of distinct keys for a fixed number of entries in the file, the better the key will be for retrieval purposes. Given two search keys which are more or less equally specific, the one which is simpler to use is preferable.

The peculiarity of corporate-entry keys can be observed from Table 1. Even for the 8,8,8,8,8 key structure the percentage of distinct keys (33.7 percent) is low, and the maximum number of occurrences of an identical key (1304) is high. Another observation revealed by Table 1 is that as the key structure goes from five to three segments, there is a steady decrease in the percentage of distinct keys and consequently an increase in the maximum number of entries per key. However, a reduction in the number of characters in a segment does not cause a great deal of deterioration. For example, for 8,8,8,8,8 keys, the percentage of unique keys and the maximum number of entries per key are respectively 33.7 percent and 1304, while for 2,2,2,2,2 keys, the corresponding figures are 32.3 percent and 1307.

Thus, the 2,2,2,2,2 key structure seemed a good candidate for a corporate

Table 1. Number of Distinct Keys and Maximum Number of Identical Entries Per Key for Different Key Structures in 68,169 MARC II Records.

Key Structure	*Number of Distinct Keys*	*Number of Distinct Keys as a Percent of Total Number of Records*	*Maximum Number of Entries Per Key*
8,8,8,8,8	22982	33.7	1304
8,8,8,8,0	20476	30.0	1305
8,8,8,0,0	16283	23.9	1802
4,2,2,2,2	22411	32.9	1307
4,2,2,2,1	22120	32.4	1308
4,2,2,2,0	19513	28.6	1311
4,2,2,1,0	18589	27.3	1311
4,2,2,0,0	14801	21.7	1807
3,3,2,2,2	22417	32.9	1307
3,3,2,2,1	22132	32.5	1308
3,3,2,2,0	19560	28.7	1311
3,3,2,1,0	18654	27.4	1311
3,3,2,0,0	14922	21.9	1806
2,2,2,2,2	22053	32.3	1307
2,2,2,2,1	21743	31.9	1308
2,2,2,2,0	19034	27.9	1311
2,2,2,1,0	18036	26.5	1311
2,2,2,0,0	13842	20.3	1807
1,1,1,1,1	19028	27.9	1308

entries index and therefore the number of entries per reply for this key structure was more intensely studied.

On the average it is desirable that the number of replies per query be such that information by which the user can choose among the possible replies can be displayed on a single CRT screen. This maximizes the utility of a computer system, since it minimizes the amount of system activity to promptly satisfy a user's request. Since some query keys produce but one reply while others produce hundreds of candidate records, it is necessary to use the mathematics of probability to determine the likely long-term effect of a given choice of system parameters. Using the approach indicated

Table 2. Average Number of Entries Per Reply for Key Structure 2,2,2,2,2 for Various Multiplicity of Entries.

Maximum Frequency of Any Entries in File	*Total Records in File*	*Percent of Total Records*	*Number of Distinct Keys Eliminated*	*Average Number of Entries Per Reply*
19	44174	64.8	389	5.0
29	48127	70.6	223	6.6
39	50854	74.6	142	8.1
49	52422	76.9	107	9.1
59	53513	78.5	87	10.1

as useful by Guthrie and Slifko, the analysis of the effect of various choices of search key becomes the following.

Assume that every entry has an equal probability of being accessed. Then, in attempting to retrieve each entry once, keys having i number of entries will cause a total of i^2 entries to be accessed. If f_i denotes the frequency of keys having i number of entries and M denotes the maximum allowable occurrences of any key in the file, the average number of entries per reply $\bar{y}$, is given by:

$$\bar{y} = \frac{\sum_{i=1}^{M} i^2 f_i}{\sum_{i=1}^{M} i f_i}$$

where $\sum_{i=1}^{M} i f_i$ is the number of entries in the file whose derived keys have a frequency of M or less.

The above formula yields the average number of entries per reply for the 2,2,2,2,2 key to be much larger than 20 for $M > 100$; but some 2,2,2,2,2 keys corresponded to more than 500 file entries. A typical CRT display terminal can accommodate only ten or fewer entries per screen. Therefore, if the average number of entries per reply is desired to be ten or fewer, it is necessary either to ignore entries with high multiplicity or to adopt a different scheme of storing and retrieving such items, in which case the mathematical result would be the same as ignoring high-frequency items.

The average number of entries per reply was computed for five different values of M (19,29,39,49, and 59); the results of these computations are in Table 2, which reveals that if keys in the file are allowed a maximum recurrence of 39 entries per key, it would be possible to have keys in the main index for about 75 percent of total records, while entries for only 142 high frequency keys would have to be shunted to a secondary index. In this case, the average number of entries per reply would be about eight.

Table 3 gives the probability of number of entries per reply for the index file consisting of 50,854 (out of a total of 68,169) records with the maximum frequency of any key in the file being 39. For preparing this table the assumption is made that each entry in the file has an equal probability of being accessed. Thus the probability of obtaining i entries per reply is given by:

$$P(i) = \frac{i f_i}{\sum_{j=1}^{M} j f_j}$$

where f_i is frequency of keys occurring exactly i number of times in the index file. An inspection of this table shows that in 87.7 percent of the

Table 3. Probability of Number of Entries Per Reply for an Index File Using 2,2,2,2,2 Key.

Number of Entries	*Frequency*	*Probability Percentage*	*Cumulative Probability Percentage*
1	14820	29.1	29.1
2	2893	11.4	40.5
3	1276	7.5	48.0
4	726	5.7	53.7
5	427	4.2	57.9
6	312	3.7	61.6
7	248	3.4	65.0
8	195	3.1	68.1
9	150	2.6	70.7
10	120	2.4	73.1
11	78	1.7	74.8
12	88	2.1	76.9
13	56	1.4	78.3
14	71	1.9	80.2
15	62	1.9	82.1
16	48	1.5	83.6
17	41	1.3	84.9
18	28	1.0	85.9
19	24	0.9	86.8
20	22	0.9	87.7
21	18	0.7	88.4
22	16	0.7	89.1
23	23	1.1	90.2
24	25	1.1	91.3
25	13	0.7	92.0
26	9	0.4	92.4
27	12	0.7	93.1
28	18	1.0	94.1
29	10	0.5	94.6
30	11	0.7	95.3
31	11	0.7	96.0
32	13	0.8	96.8
33	6	0.4	97.2
34	9	0.6	97.8
35	7	0.4	98.2
36	6	0.5	98.7
37	11	0.8	99.5
38	5	0.3	99.8
39	2	0.2	100.0

time there would be 20 or fewer replies. This represents two screensful of information on a typical CRT display.

CONCLUSION

A file containing only those entries for which the frequencies of 2,2,2,2,2 search keys is 39 or fewer would produce 20 or fewer entries per reply approximately 88 percent of the time, but such a file excludes 142 high frequency keys for 17,315 of a total of 68,169 entries. Therefore, a special technique for handling corporate-entry derived keys of high multiplicity is desirable.

REFERENCES

1. A. L. Landgraf and F. G. Kilgour, "Catalog Records Retrieved by Personal Author Using Derived Search Keys," *Journal of Library Automation* 6:103-8 (June 1973).
2. F. G. Kilgour, P. L. Long, and 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* 7:79-82 (1970).
3. F. G. Kilgour, P. L. Long, E. B. Leiderman, and A. L. Landgraf, "Title-Only Entries Retrieved by the Use of Truncated Search Keys," *Journal of Library Automation* 4:207-10 (Dec. 1971).
4. P. L. Long and F. G. Kilgour, "A Truncated Search Key Title Index," *Journal of Library Automation* 5:17-20 (March 1972).
5. Kilgour, Long, Leiderman, "Retrieval of Bibliographic Entries."
6. G. D. Guthrie and S. D. Slifko, "Analysis of Search Key Retrieval on a Large Bibliographic File," *Journal of Library Automation* 5:96-100 (June 1972).
7. Landgraf and Kilgour, "Catalog Records Retrieved."