A TRUNCATED SEARCH KEY TITLE INDEX

Philip L. LONG: Head, Automated Systems Research and Development and Frederick G. KILGOUR: Director, Ohio College Library Center, Columbus, Ohio.

An experiment showing that 3, 1, 1, 1 search keys derived from titles are sufficiently specific to be an efficient computerized, interactive index to a file of 135,938 MARC II records.

This paper reports the findings of an experiment undertaken to design a title index to entries in the Ohio College Library Center's on-line shared cataloging system. Several large libraries participating in the center requested a title index because experience in those libraries had shown that the staff could locate entries in files more readily by title than by author and title. Users of large author-title catalogs have long been aware of great difficulties in finding entries in such catalogs. Since the center's computer program for producing an author-title index could be readily adapted to produce a title index, it was decided to add title access to the system.

A previous paper has shown that truncated three-letter search keys derived from the first two words of a title are less specific than author-title keys (1). Earlier work had revealed that addition of only the first letter of another word in a title improved specificity (2). Therefore, the experiment was designed to test the specificity of keys consisting of the first three characters of the first non-English-article word of the title plus the first letter of a variable number of consecutive words.

The experiment was also designed to produce an index that catalogers could use efficiently and that would operate efficiently in the computer system. It was assumed that the terminal user would have in hand the volume for which an entry was to be sought in the on-line catalog. The index was not to be designed for use by library users; subsequent experiments will be done to design an index for nonlibrarian users.

Other investigations into computerized, derived-key title indexes include the previous paper in this series to which reference has already been made (1) and development of a title index in Stanford's BALLOTS system (3). Although Stanford has not published results observed from experiment or experience that describe the retrieval specificity of its technique, it is clear that the Stanford procedure is not only more powerful than the one described in this paper but also more adaptable for user employment. The Stanford index is probably less efficient.

MATERIALS AND METHODS

A file of 135,938 MARC II records was used in this experiment. This file contains title-only and name-title entries, and keys were derived from titles in both types of entries. A key was extracted consisting of the first three characters of the first non-English-article word of each title plus the first character of each following word up to four. If there were fewer than four additional words, the key was left-justified, with trailing blank fill. Only alphabetic and numeric characters were used in key derivation; alphabetic characters were forced to uppercase. All other characters were eliminated and the space occupied by an eliminated character was closed up before the key was derived. A total of 115,623 distinct keys was derived from the 135,938 entries.

These 115,623 keys were then sorted. Each key in the file was compared with the subsequent key or keys and equal comparisons were counted. A frequency distribution by identical keys was thus prepared, and a table constructed of percentages of numbers of equal comparisons based on the total number of distinct keys. This table contains the percentage of time for expected numbers of replies based on the assumption that each key had a probable use equal to all other keys.

Next, by eliminating the fourth single character and then the fourth and third, files of 3,1,1,1 and 3,1,1 keys were prepared from the 3,1,1,1,1 file. For example, the 3,1,1,1,1 key for Raymond Irwin's *The Heritage of the English Library* is HER, O, T, E, L; the 3,1,1,1 key for this title is HER, O, T, E; and the 3,1,1 key, HER, O, T. The same processing given to the 3,1,1,1,1 file was employed on these two files.

RESULTS

Table 1 contains the maximum number of entries in 99 percent of replies. Inspection of the table reveals that there is a large increase in specificity when the key is enlarged from 3,1,1 to 3,1,1,1; the maximum number of entries (99+ percent of the time) drops from twelve to five. However, when the key goes to 3,1,1,1,1, the number of entries per reply goes down only to four from five.

The percentage of replies that contained a single entry was 67.8 for the 3,1,1 key, 84.0 for the 3,1,1,1 key, and 90.0 for the 3,1,1,1,1 key.

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

Search Key	*Title Index Entries*	
	Maximum Entries Per Reply	*Percentage of Time*
3, 1, 1	12	99.0
3, 1, 1, 1	5	99.1
3, 1, 1, 1, 1	4	99.2

The Irascope cathode ray tube terminals used in the OCLC system can display nine truncated entries on the one screen, and it is felt that catalogers can use with ease up to two screensful of entries. Therefore, the keys producing more than eighteen titles were listed. For 3,1,1,1,1 there were only 33; for 3,1,1,1 there were 67; and for 3,1,1 there were 357.

The maximum number of identical keys was 321 for 3,1,1,1,1 and 3,1,1,1; the key was PRO, b̸, b̸, b̸, b̸, most of which was derived from "Proceedings." For 3,1,1 the maximum was 417, for HIS, O, T—"History of the."

DISCUSSION

It is clear from the findings that a 3,1,1 search key is not sufficiently specific to operate efficiently as a title index in a large file. However, the 3,1,1,1 key appears to be sufficiently specific for efficient operation, while the 3,1,1,1,1 key does not appear to possess sufficient increased specificity to justify its additional complexity.

The observation that there is a large increase in specificity between keys employing three- and four-title words that constitute Markov strings suggests that the second and third words may be highly correlated. Indeed this suggestion is substantiated by the maximum case for 3,1,1—HIS, O, T. In the more-than-eighteen group for 3,1,1,1, these characters occurred in seven keys for a total of 206 entries, and for 3,1,1,1,1 they did not occur at all in the more-than-eighteen group.

CONCLUSION

This experiment has shown that a 3,1,1,1 or 3,1,1,1,1 derived search key is sufficiently specific to operate efficiently as a title index to a file of 135,938 MARC II records. Since a previous paper observed that as a file of entries increases the number of entries per reply does not increase in a one-to-one ratio (1), it is likely that these keys will operate efficiently for files of considerably greater size.

REFERENCES

1. Frederick G. Kilgour, Philip L. Long, Eugene B. Leiderman, and Alan L. Landgraff, "Title-Only Entries Retrieved by Use of Truncated Search Keys," *Journal of Library Automation* 4:207–10 (Dec. 1971).
2. Frederick G. Kilgour, "Retrieval of Single Entries from a Computerized Library Catalog File," *Proceedings of the American Society for Information Science* 5:133–36 (1968).
3. Edwin B. Parker, *SPIRES (Stanford Physics Information REtrieval System) 1969–70 Annual Report* (Palo Alto: Stanford University, June 1970), p. 77–78.