ANALYSIS OF SEARCH KEY RETRIEVAL ON A LARGE BIBLIOGRAPHIC FILE

Gerry D. GUTHRIE, Steven D. SLIFKO: Research & Development Division, The Ohio State University Libraries, Columbus, Ohio

Two search keys (4,5 and 3,3) are analyzed using a probability formula on a bibliographic file of 857,725 records. Assuming random requests by record permits the creation of a predictive model which more closely approximates the actual behavior of a search and retrieval system as determined by a usage survey.

INTRODUCTION

Systems planners are hard pressed to accurately predict the access characteristics of search keys on large on-line bibliographic files when so little is known about user requests. This paper presents a realistic model for analyzing different search keys and, in addition, the results are compared to actual request data gathered from a usage survey of the Ohio State University Libraries Circulation System.

A number of papers are available in the literature concerning search key effectiveness; however, all of these were done on relatively small data bases (1-5). Of particular importance to this paper is Kilgour's article on truncated search keys (6).

PURPOSE

The purposes of this study are (1) to determine the comparative effectiveness of the 4,5 and 3,3 search keys, (2) to compare two predictive models, and (3) to test the results with an actual usage survey.

METHOD

The Ohio State University Libraries Circulation System contained at the time of this study 857,725 titles representing over 2.6 million volumes in the

OSU collection. The data base used for this study was the search key index file which contained one search key for each title in the master file.

The search key is composed of the first four letters of the author's last name and the first five letters of the first word of the title excluding non-significant words (4,5 key). Title words are passed against a stop-list to determine significance. The stop-list contains the words: a, an, and, annual, bulletin, conference, in, international, introduction, journal, of, on, proceedings, report, reports, the, to, yearbook. The search key file is in sequence by search key.

For comparative purposes, a second search key file was created and sorted which contained a 3,3 key (the first three characters of the author's last name and the first three characters of the first significant word of the title.)

The two files of sorted search keys were then processed by a statistical analysis computer program. This program created a frequency distribution table of identical keys, i.e., how many keys were unique, duplicated once, duplicated twice, etc. From this table two models were compared.

Model 1:

File entry was viewed as a random process with choice of any unique search key equiprobable. This model has been suggested in the literature mentioned earlier. It states that if x_i number of keys will return i matches then the probability of a file search returning i matches may be written:

$$P(i) = x_i/K_u$$

where K_u is the total number of unique file keys.

Likewise, the cumulative probability for I or fewer matches is

$$P(I) = \sum_{i=1}^{I} P(i) = \left(\sum_{i=1}^{I} x_i \right)/K_u$$

Model 2:

File entry is viewed as a random process with the choice of any record equiprobable. Thus,

$$P(i) = ix_i/R_t$$

where R_t is the total number of file records. Correspondingly,

$$P(I) = \sum_{i=1}^{I} P(i) = \left(\sum_{i=1}^{I} ix_i \right)/R_t$$

Survey:

The Ohio State University Libraries Automated Circulation System includes a telephone center to which patrons may telephone requests for

library holdings information and for checking out and renewing books. Telephone operators, sitting at cathode ray tube (CRT) terminals, translate the patron's author-title request into a 4,5 search key and proceed with a file search. By having the telephone operators treat telephone calls as random input to the system and recording the number of matches returned for each search used, results can be generated in the same form that both of the models take, i.e., I or fewer matches have been returned $P(I) \times 100$ percent of the time.

This is a relatively easy survey to conduct since the output list of matching records for any particular key entry is headed with the exact number of matches which follow. The sample size was 1000 information requests recorded over two one-week periods separated by one month. Before these two subsamples were merged, statistical analysis on their individual means (for percent of 10 or fewer matches) signified they were identical at the 99 percent confidence level.

RESULTS

The results predicted by the two models for both a 4,5 and 3,3 search key for 1-10 matches appear in Tables 1 and 2.

The figures pertaining to the 4,5 key can be compared directly to the data received from the survey conducted through the OSU Library's telephone center. This comparison is shown in Table 1 for 1-10 matches.

Table 1. File Access Comparisons (4,5 search key).
(Percent of time I or fewer matches returned)

I	*Actual Survey*	*Model 1 (random key)*	*Model 2 (random record)*
1	35.9	81.3	55.7
2	53.8	92.9	71.6
3	66.0	96.3	78.5
4	73.1	97.7	82.4
5	78.5	98.4	84.9
6	81.3	98.8	86.6
7	83.8	99.1	87.8
8	85.6	99.3	88.8
9	86.6	99.4	89.6
10	87.8	99.5	90.2

To acquire a 99 percent upper confidence limit on the percent of requests returning 10 or fewer matches, the normal distribution was used as an approximation to the binomial distribution ($n = 1000$, $p = .878$) producing an upper limit of 90.2 percent.

Table 2. File Access Comparisons (3,3 search key).
(Percent of time *I* or fewer matches were returned)

I	*Model 1 (random key)*	*Model 2 (random record)*
1	64.3	28.0
2	81.0	42.5
3	87.9	51.7
4	91.6	58.0
5	93.7	62.7
6	95.1	66.3
7	96.1	69.3
8	96.8	71.8
9	97.3	73.9
10	97.7	75.7

DISCUSSION

In Table 1 the results of the survey show that 87.8 percent of all searches recorded returned 10 or fewer titles. In Model 1, assuming that requests of the file are random with respect to search key, it is predicted that 99.5 percent of all searches will return 10 or fewer titles. All predicted percentages for Model 1 are consistently higher than observed results.

The predicted response in Model 2 more closely approximates the observed behavior of the system as the number of responses increases. However, Model 2 is also consistently higher than the actual survey. Comparing Model 1 and Model 2 only, it is apparent that assuming a random record request more accurately reflects the true usage of a library collection.

The lower percentages recorded in the actual survey may be attributable to a number of variables not taken into consideration in this study. Clustering due to common English word titles and common names may account for the greater part of this difference.

Table 2 shows the results of predicted response for a 3,3 search key. In this table, Model 2 predicts that only 75.7 percent of requests will return 10 or fewer titles. Equally important, only 28.0 percent of the requests will return a single record.

CONCLUSION

In predicting the expected behavior of an information retrieval system, it is more accurate to assume random requests by record than to assume random requests by search key. Probability predictions are deceptively high for assumed random key requests and do not reflect actual usage of the file.

Even assuming random requests by record will produce higher-than-observed results. Data calculated using Model 2 should be considered as an upper limit or "ideal" performance indicator. Regarding the results of

the random record model as the upper limit on effectiveness of the search key, the data gathered indicate that, as the search key is shortened from 4,5 to 3,3, the deviation between the random key and random record models is considerably heightened.

The 4,5 search key is more efficient for retrieval of 10 or fewer records from a large file than the 3,3 key (90.2 - 75.7 percent). Based on these data, the OSU Libraries decided to retain the 4,5 search key and not reduce it to 3,3.

Additional studies should be undertaken to determine the effects of common word usage, common names, and their relation to book usage. Secondly, the data presented here could be systematically and randomly reduced in size to predict the behavior of various search key combinations on varying file sizes.

REFERENCES

1. Philip L. Long and Frederick G. Kilgour, "A Truncated Search Key Title Index," *Journal of Library Automation* 5:17-20 (Mar. 1972).
2. Frederick G. Kilgour, Philip L. Long, Eugene B. Leiderman, and Alan L. Landgraf, "Title-Only Entries Retrieved by Use of Truncated Search Keys," *Journal of Library Automation* 4:207-10 (Dec. 1971).
3. 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).
4. Frederick H. Ruecking, Jr., "Bibliographic Retrieval from Bibliographic Input; The Hypothesis and Construction of a Test," *Journal of Library Automation* 1:227-38 (Dec. 1968).
5. William L. Newman and Edwin J. Buchinski, "Entry/Title Compression Code Access to Machine Readable Bibliographic Files," *Journal of Library Automation* 4:72-85 (June, 1971).
6. Frederick G. Kilgour, Philip L. Long, and Eugene 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* 7:79-81 (1970).