A Possibly Inexpensive Attachment for a Microfilm Reader to Permit Synchronized Coordinate Searching[1,2,3]

By JOHN O'CONNOR
Institute for Cooperative Research,
University of Pennsylvania, Philadelphia, Penna.
Received June 6, 1961

1. General Function of the Device.—Suppose we have a set of texts or abstracts on microfilm with each occupying one frame. And suppose that each frame has been coordinately indexed though the index sets are not necessarily on the microfilm.

The kind of device to be described will permit coordinate searching of a reel of the microfilm at anywhere from a few hundred to perhaps over a thousand frames per minute, depending on the particular arrangement used and certain characteristics of the index sets. The coded index sets are on an auxiliary record which will be described. Only a search question which is a conjunction of terms (and some negated terms, if desired) can be handled in a one-pass search. Whenever a search question is satisfied by an index set, the corresponding text frame is on the microfilm viewing screen. Since the searching of index sets is done by human sensing, the searching is subject to fatigue limits and possible "careless errors." The cost of purchasing and assembling the materials needed, other than the microfilm viewer of course, might be relatively small.

2. Physical Description.—Consider a scroll of paper between two rollers A and B, with A directly above B at a distance of a foot or two. If a crank is appropriately geared to the axles of A and B, turning the crank one way will unroll the scroll from A down to B, and turning it the other way will do the reverse. Suppose that the crank is geared to a microfilm reel in a viewer standing beside the scroll, so that when the scroll moves downward from A to B, the reel moves forward, and similarly for the opposite directions. An inexpensive electric motor might power these movements, although this is not essential. An apparatus such as this is the basic form of auxiliary record on which index sets are to be coded for searching.

3. Scan Column Coding.—The index sets will be represented on the scroll, for human visual searching as the scroll unrolls from A to B. How can the index sets be coded to facilitate such searching?

In view of the gearing to the microfilm reel which I have already described, a single row (width presently unspecified) across the entire scroll should be devoted to each frame's index set. Then how should the index terms assigned to that frame be coded in the row?

Suppose there are a total of V terms in the indexing vocabulary. We might divide the scroll into V different columns, and assign each term to a different column. Thus if the index set for frame 2862 includes the term *chlorine*, for example, then on the row for frame 2862, in the column for *chlorine*, a simple mark such as a dash is entered. I'll call each column in such an arrangement a "single-term scan column."

If for any reason V columns are too many, modification of the above representation can be made. To each column, a number of different terms are assigned. Each of the terms represented in a column is then coded by a different mark. For example, *chlorine* might be represented in its column by b, and if *radiation* and *sulfur* are two of the other terms represented in the same column they might be coded by c and d, respectively. I'll call each column in an arrangement of this kind a "multi-term scan column," or usually just a "scan column."[4]

It is evident how a conjunction search would be made on a scroll with single-term scan columns. For example if the search terms were *chlorine* and *insecticide*, the *chlorine* column could be scanned for marks. Each time a mark was found, the insecticide column in the same row would be looked at. If the scan-columns were multi-term rather than single term, the character-abbreviation for each term (*e.g.*, b for *chlorine*) would have to be looked up for each search term, and then the column to which chlorine and some other terms had been assigned scanned for occurrences of b.

Suppose two terms are coded by different marks in the same scan column. If a particular index set contains both those terms ("multiple entry"), then how should the terms be coded in the row for that index set? If the rows are wide enough, or the columns are, both marks can be entered in the same row and column. Otherwise one mark can be placed in the first empty row above. Special marks, *e.g.*, primes or under-and-over-linings, should then accompany both marks to indicate the displacement. If no two terms assigned to a scan column are strongly positively correlated, then the frequency of such entries is a function of the column density, d. The probability of a row having a multiple entry in a column is $(1/2)d^2$. Further details are in "The Scan Column Index."[2]

(1) Research sponsored by the Information Systems Branch, Office of Naval Research [Contract Nonr 551 (35)], and by the Air Force Office of Scientific Research Reproduction in whole or in part permitted for any purpose of the United States Government.

(2) The basic idea of the present paper we suggested briefly on page 51 of "The Scan Column Index," a report by the author, which may still be avialable from Contracts, Remington Rand Univac, Bluebell, Pa. An abridged version of the report appeared in *American Documentation*, April, 1962.

(3) A number of people contributed ideas to the notion of a scan column index. The scroll idea, which is central in the present paper, is due to John Mauchly. Claire Schultz also made some suggestions.

(4) Of course a scroll might have some single-term and some multi-term scan columns.

3A. Scan Columns on Microfilm.—Two reviewers of this paper suggested that the scan columns might also be on microfilm, either along the edge of the text film, or on a separate reel in a second (sychronized) viewer.

If scan columns are on the text film, there will be room for relatively few of them. This means that each column will be relatively dense and therefore searching relatively slow (see section 5, second paragraph, and section 12, first paragraph).

If the scan columns are on a separate film, there will be room for more of them, but apparently not nearly so many as on a scroll (see section 12, second paragraph). There might be cases where it is best to use either of these microfilm forms of scan column. However, the remainder of the present paper will discuss non-microfilm scan column scrolls.

4. Synchronization and Scroll Marking.—Suppose a row indicates that a particular microfilm frame is relevant to the search question. Where will that row be when the corresponding frame is in the viewer? Either immediately beside the scroll or across its face should be a horizontal marker, such as a thin rod. The reel-scroll synchronization should be such that when a row reaches this "row marker," the corresponding frame is in the viewer.

In a book form scan column index it does not appear to be necessary to mark every row and every column with a separate line. A vertical line drawn after every third column and a horizontal line after every fourth row (for example) seem to work satisfactorily. A similar marking of rows and scan columns on the scroll also seems likely to be sufficient. The figure illustrates how part of a scan column scroll with column density of about $\frac{1}{20}$ might appear. The numbers on the left are document numbers.

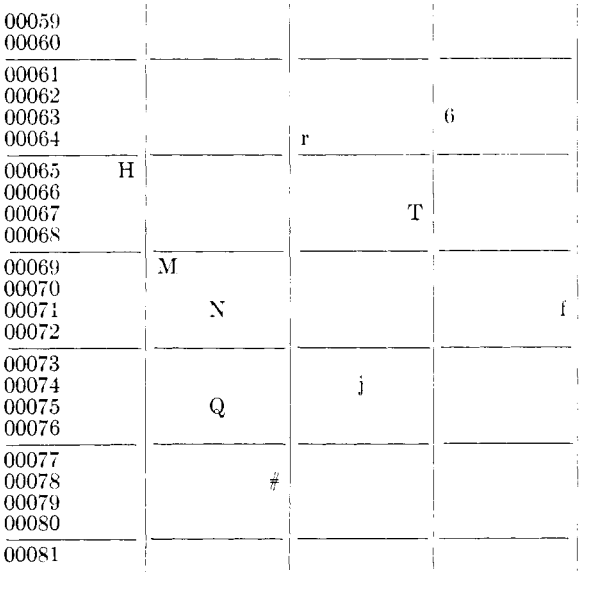

Figure I.

5. Performance Characteristics.—What about search speed, fatigue, "careless error," wear, accuracy of synchronization between scroll and microfilm, and initial and operating costs of a scan column scroll? No model of the device has been built, so there is no working experience to provide evidence. However a few remarks can be made in this and the next several sections.

One might expect that column density (frequency of marks of all kinds in a column) makes a great difference to single column search speed. A few trials with sample pages of a book-form (multi-term) scan column index provide some confirmation of this. The trials suggested that for column densities ranging from $\frac{1}{5}$ to $\frac{1}{18}$, scanning speed ranged from about five hundred to one thousand documents per minute. A similar few trials suggested that looking at a second column in a row required from two to five seconds.

In the case of a scroll, pages don't have to be turned, as in a book form scan column index, and this might help searching speed some. On the other hand, if a very wide scroll is used, examining a second column might sometimes require significantly more than five seconds. I will say more about search speed later.

6. Fatigue and Careless Errors.—Concerning both fatigue and "careless errors" in scanning, my guess, but it is only that, is that for columns of low density, say $\frac{1}{10}$ or less, these problems might not be more significant, perhaps less so, than for other searching systems which primarily use human sensing.

If a person with some interest in the subject matter is doing the scanning, the examinations of relevant frames in the viewer, when they are found, constitute breaks in the scanning routine, and rewards of searching, which might reduce further the chance of fatigue and error. This kind of factor also helps some other search systems which primarily use human-sensing.

7. Wear and Synchronization.—The questions of wear and of reel-scroll synchronization are more strictly "hardware" problems rather than human behavior questions. What kind of material to use for a scroll is obviously not a trivial question. The scroll should not tear, yet it should turn easily, and be easy to mark, and the material should not be unduly expensive. Further, tearing is a function of the mechanics of the rollers' operation as well as of the scroll material. Only some actual trials would give definite information on these matters.

The synchromization problem is not just a matter of gearing, but a question of possible scroll slippage. It hardly seems insurmountable, but it would require some attention in an actual model. The synchronization required is not an all-or-nothing matter, but a question of degree. For instance, suppose that when a row reaches the horizontal marker the corresponding frame is mostly in the viewer. Then if hand positioning is convenient, this should be satisfactory for many purposes. And if each frame has a serial number, and each row has a corresponding serial number on the side nearer the microfilm viewer, then even less synchronization than this might be satisfactory for many purposes.

8. Possible Costs.—Cost figures for the device I have been describing are (exclusive of microfilm viewer) unknown. But my impression is that the purchase and appropriate assembly of the material involved should be relatively inexpensive. The more precise the reel-scroll synchronization required, probably the more expensive is

(5) There is a procedure for assigning terms to columns so that all multi-term scan columns are of about the same density; see section 2-14 of the report, "The Scan Column Index."

the device to make, though perhaps still within a modest upper bound.

The labor cost of entering index sets looks as though it should be comparable to that for other systems, though once more no evidence from operating experience is available.

9. Further on Search Speeds.—The rate at which index sets can be searched depends on single column scanning speed and the time required for looking at other columns when relevant. For example, suppose a single column can be scanned at one thousand rows per minute, and a reference to a second column requires five seconds. Suppose also that in the first column one row in a hundred contains the term being scanned for. Then every thousand rows require ten second column references, which cost fifty seconds. Thus one thousand rows are examined in almost two minutes, and the actual searching speed is about five hundred rows per minute.

If the first two terms searched for are strongly positively correlated (*i.e.*, the presence of either strongly increases the probability of occurrences of the other), and if there is a third search term, then looking at the third column will further add significantly to search time. However, if the first two search terms are not strongly correlated, and not both very frequent, then third column references will add little to search time. I will consider only the second column reference here.

In general, suppose the rate of single column scanning is s rows per second, the frequency of the term looked for in the column is f (*e.g.*, $f = 1/100$ means an average of one occurrence per one hundred rows), and the time in seconds for looking in a second column is h. If t seconds are spent in single column scanning, then ts rows will be covered. However the search will also require tsf second column references, which will cost a total of $tsfh$ seconds. Thus the total search time for ts rows is $t + tsfh$. The actual search rate in rows per second is then $ts/(t + tsfh) = s/(1 + sfh)$ rows per second. For example if s = one thousand rows per minute, which is about seventeen rows per second, $f = 1/100$, and h is 5, then the actual search rate is nine rows per second, *i.e.*, five hundred and forty rows per minute.

10. Some Details on Search Speeds.—The formula for the actual search rate can be rewritten as $1/(1/s + fh)$. This shows that while one can always increase the actual search rate by making s larger, the upper limit on such an increase is $1/fh$. For example, if $f = 1/100$ and h = five seconds, the search rate cannot exceed twelve hundred rows/min.. This limiting figure represents the case of *all* search time being required for references to the second column.

Suppose we have a search rate $s/(1 + sfh)$, and it is possible, for that fh, to multiply the search rate by M, if we multiply the single column scanning rate s by Z. Then we set $Ms/(1 + sfh) = Zs/(1 + Zsfh)$. Some algebra gives $Z = M/[1 - (M - 1)sfh]$, which is meaningful here only when both M and Z are positive. For example, if $fh = 1/20$, then the maximum search rate possible by increasing s is $1/fh = 20$ rows/sec.. For $s = 17$ rows/sec., the actual search rate is 9 rows/sec.. Therefore for an initial $s = 17$ we can seek to double the search rate ($M = 2$) by multiplying the single column scanning rate by Z. However, the formula for Z shows that to double the actual search rate in this case we would have to multiply the single column scanning rate by more than thirteen (for $M = 2$, $Z = 13.3$).

Quite different figures result for $s = 17$ rows/sec. if $fh = 1/100$ (*e.g.*, $f = 1/100$ and h = one second). Then the actual search rate is 14 rows/sec. And to double the actual search rate requires only multiplying s by 2.4 (for $M = 2$, $Z = 2.4$).

11. Design Features to Increase Search Speeds.—f can be influenced by the searcher in any particular case by selecting the term with lowest frequency for first column scanning.

h can perhaps be influenced significantly by the searcher, if there are more than two search terms, by selecting as the first two terms for search those in two scan columns which are near one another, if this is possible. However both of such terms might have high frequencies, or they might be strongly positively correlated with each other, so that they would not be good choices.

The most important question is, how can the general design of the scroll device reduce h. h might be reduced by a movable vertical wire, attached to the horizontal row marker, which could be slid to the scan column which was the second column for a particular search. Another sliding column marker might also help in quickly returning to the first column to resume scanning after a second column reference. A reviewer suggested a perforated adjustable mask to expose only the columns being scanned.

When looking in a second column on a row, one is looking for a particular symbol, *e.g.*, a C. But a C will be fairly infrequent, or perhaps quite infrequent (*e.g.*, one C per one hundred rows) in the second column. Therefore it is usually unnecessary to be sure that the second column reference is to the correct row. Only in the infrequent cases when a C is found must this be checked. Similarly, it is not necessary to look in exactly the right column for a C; only when a C is found is exactness necessary. Thus if the wire column markers suggested in the preceding paragraph were used, they would not need to be precisely vertically synchronized with the search columns which they would help to locate. It would be sufficient if they were located so as not to obscure or exclude the scan columns of interest.

It is relevant to remark that in most of the trials which suggested the figure of two to five seconds for a second column reference (section 5), the second columns were quite crowded, with densities of 1/4 or more. A scan column scroll with average column densities of 1/10 or less might, even aside from the suggestions of the preceding two paragraphs, require much less time for second column references. Once more, however, there is no evidence concerning this from actual operation.

12. Column Density and Search Speeds.—If the average number of terms per index set is t, and there are c scan columns, then the average scan column density is t/c. Thus the more scan columns, the less density per column.[5] Of course this only holds for multiterm scan columns.

A scroll can hold many columns. For example, if each scan column is an eighth of an inch wide, a three foot wide scroll can contain almost three hundred columns. If there are an average of fifteen terms per index set, this means an average column density of about 1/20. This might mean an average rate for single column scanning of one thousand frames or more per minute (see section 3). A still wider scroll would cut density even more (except for single terms of five percent frequency).[6]

13. Wide Scrolls and Divided Scrolls.—A way of handling the wide scroll problem is to divide the total scan column scroll. For instance, the first six inches of scan columns columns on another, *etc.* When a search is to be made, then the particular narrow scrolls containing the relevant scan columns would be selected and mounted on the rollers.

Such a divided scroll might present problems of synchronization between different narrow scrolls because of differing slippages. However the next to the last paragraph of section 11 is also relevant here.

A divided scroll complicates entry of new index sets. Perhaps such entry can best be handled by lining up all the narrow scrolls on a pair of long rollers (as though the total scan columns' scroll had never been divided) and thus reducing the procedure of entering new index sets to something like its form for an undivided scan column scroll. However imperfect synchronization among the narrow scroll might require significantly more time for entering new index sets.

The narrow scrolls selected for a search also have to be mounted on one roller and then attached to the other, while a total scroll which is not divided can remain always on both rollers.

In general, the conflicting values involved in a divided scroll would have to be balanced against each other in terms of the needs of the particular application.

14. Multi-Colored Marks.—Another possible way to increase speed of scanning a column is to use marks of several different colors in a column. For example, if someone is looking for a red d in a column, marks of other colors in the column might slow him down almost as little as would empty space.

A simple way of getting something of this effect would be to use boldface, regular, and lightface characters.

Another possible technique is the use of radically different shaped characters, *e.g.*, (a), (b), (c), d, e, f. If someone was scanning for (b), then d, e, and f might slow him little more than white space would.

15. Motorized Scroll Turning.—Suppose through use of a very wide scroll, a divided scroll, or varicolored marks, we have reduced column density very greatly, for example to $\frac{1}{50}$. If the scroll turns quickly enough, we might conjecture that single column scanning speed could be quite rapid, perhaps several thousand rows per minute. An inexpensive electric motor might be able to turn the scroll quickly enough.

16. Possible Use of Document Grouping.—Suppose that by means of a very wide scroll, a divided scroll, or varicolored marks, we have reduced column density greatly, once more for example to $\frac{1}{50}$. Suppose then that we record in each row not one index set but three.[7] Then the average column density will be about $\frac{1}{17}$, which might give a single column scanning rate of about one thousand rows per minute (section 5). But there will be only one third as many rows to be searched as there are index sets in the collection, so that in effect the single column scanning rate is about three thousand rows per minute.

There is a possibility of false drops resulting from document grouping. For example, if index set 4056 has *chlorine* but not *insecticide*, index set 4057 the reverse, and they are both represented in row 1352, then that row would be incorrectly selected as relevant for the search question *chlorine* and *insecticide*. However the number of such false drops can sometimes be very small.[8] Further, special marks can be used to distinguish the various index sets in a group. For example b, (b), and *b* might represent *chlorine* included in the index sets 4056, 4057, and 4058 respectively.

When grouping is used, in the formulas given in section 9, s is replaced throughout by gs. For gs is the single column rate of searching index sets, and it is this rate which determines the number of second column references required.

If, as a result of grouping, a row represents the index sets for (*e.g.*) three frames, than reel-scroll synchronization is a bit complicated. However the following might be satisfactory. When a row is at the row marker (section 3, near end), the middle frame of the three frames represented in the row is in the viewer. Variations are obvious for cases when other than three index sets are represented in each row.

(6) If there are V terms, and an average of t terms per index set, then the average term frequency is t/V. For example, if $t = 15$ and $V = 1{,}000$, then the average term frequency is 1.5 per cent.

(7) This is an application of "document grouping." See "The Possibilities of Document Grouping" in "Information Retrieval and Machine Translation," Part I, edited by Allen Kent, Interscience Publishers, Inc., New York, N. Y., 1960.

(8) If p_x and p_y are the frequencies of terms X and Y, there are g index sets per group, and index sets are assigned randomly to groups, then for the search X and Y the probable number of false drops is about $(g - 1)p_x p_y R$, where R is the total number of index sets.