XXII. DISTINGUISHING RETRIEVED FROM NONRETRIEVED INFORMATION: THE CUT-OFF PROBLEM

S. J. Sillers

1. The Cut-off Problem

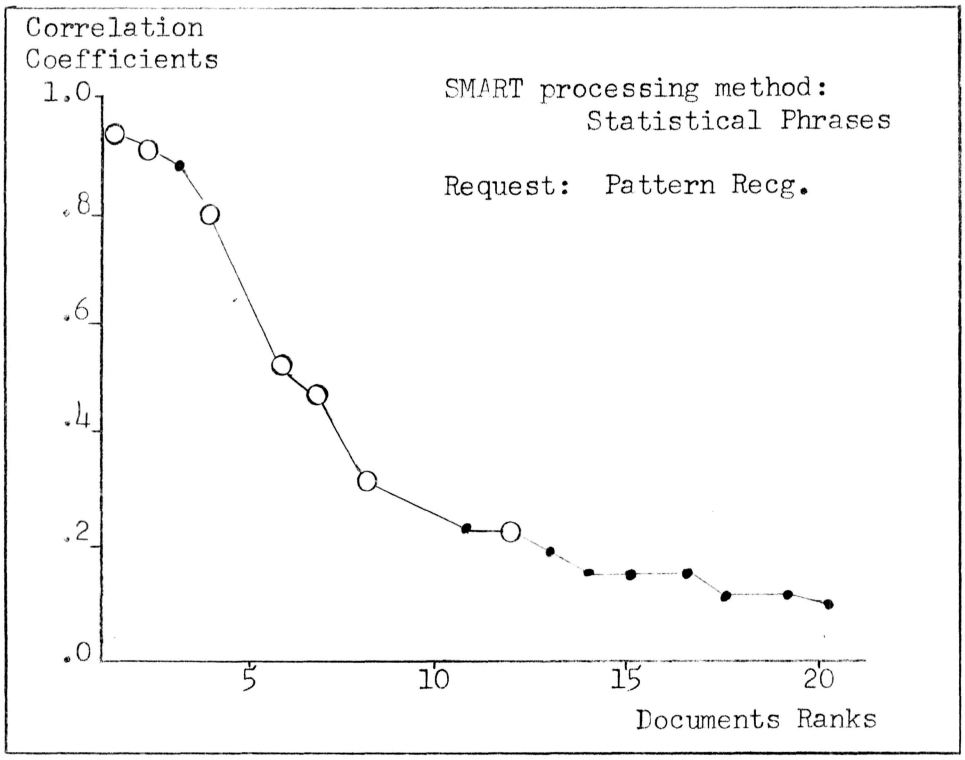
A fully automatic information retrieval system must have the capacity to recognize those documents which are relevant to a given search request. Although the final decision to accept or to reject a document is a binary one, relevant documents may differ greatly in their pertinence to the request under consideration. Moreover, since relevance is as yet not mathematically defined, there is generally no clear dividing line between the relevant and the irrelevant. Therefore, in a retrieval system a measure is usually employed to quantify the similarity between documents and requests, and a cut-off point is chosen to separate accepted (that is, hopefully, relevant) and rejected (that is, hopefully, irrelevant) documents. It is the problem of choosing this cut-off point which is considered here.

2. Possible Methods of Solution

In the SMART automatic document retrieval system a correlation between each given request and each of the documents on file is calculated after content analysis, by one of a number of possible methods. The documents are then ranked in order of decreasing correlation coefficients.

The cut-off point has heretofore been chosen by either eliminating all documents with correlation coefficients below a certain threshold value, or by simply recovering all documents in their ranked order. The first method appears unsatisfactory for use with the SMART system because of the heavy dependence of the correlation coefficients on the processing technique, while the second is practicable only under experimental conditions because of the potentially large size of the file.

A more natural cut-off might be devised through consideration of the behavior of the correlation coefficients as a function of the ranks of the documents. Although these ranks are discrete and integral-valued, the correlation points may be plotted, and continuous curve fitted through adjacent points with straight lines, for instance. The curve so obtained is monotonically decreasing, and has a slope defined everywhere but at the sample points themselves. Standard differencing methods may be used to study the properties of this curve, and hopefully, to give some insight into a solution of the cut-off problem. Figure 1 shows the first 20 points of a sample correlation curve. The circled correlation values correspond to documents actually believed relevant.



Ranked Correlation Coefficients

Figure 1

If it is assumed that the break between relevant and nonrelevant documents manifests itself by some peculiar behavior of the correlation values, several possible candidates for a cut-off point readily present themselves. The largest drop between successive correlation coefficients occurs at the point of the minimum (largest negative) first forward difference, while the most radical change in the structure of the curve lies near the point of maximum magnitude of the second forward differences. If a correlation coefficient is considered as the probability that the given document is relevant, then, since each document is evaluated independently, the product of the N highest correlation values represents the probability that all of the N corresponding documents are relevant.

In this case a confidence limit must be chosen before the cut-off point can be determined.

3. The Computer Program

To investigate the relative merits of the cut-off methods suggested above, as well as several variants of these methods, a series of cumputations were made using data obtained in previous experiments employing the SMART system. These computations were performed on the 7094 computer with a FORTRAN program designed to allow as much flexibility as possible in the combinations of cut-off methods studied during a particular run. The main routine of this program handles all input and output, as well as the computation of evaluation statistics for the results. The cut-off indices themselves were calculated by a subroutine coded in such a way that additional methods might easily be inserted. The methods included in this subroutine were as follows:

(the cut-off point, or last ordered document retrieved, is that document having the property described)

1. lowest correlation coefficient above a specified threshold;
2. last forward difference with magnitude above a specified threshold;
3. first forward difference of maximum magnitude;
4. second forward difference of largest magnitude if this difference is positive; next point if negative;
5. last second difference with magnitude above a specified threshold if this difference is positive; next point if negative;

6. last product of correlation coefficients above a specified threshold; (this product is taken over the ordered coefficients through the point under consideration).

The data required by the program consist mainly of the ordered correlation values corresponding to each request, for each of the SMART procsing methods considered. These correlation coefficients, along with the indicies of their associated documents within the document file, were entered on punched cards. For the evaluation of the cut-off methods, the indicies of those documents actually deemed relevant to each request (previously determined manually) were also supplied on cards. The remainder of the data included parameters to be used in the designation of the chosen cut-off methods.

The statistics used in judging the methods were the usual recall and precision measures, recall being defined as the proportion of relevant material actually retrieved, and precision as the proportion of retrieved material actually relevant. More precisely,

$$\text{recall} = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents in file}}$$

$$\text{precision} = \frac{\text{number of relevant documents retrieved}}{\text{total number of documents retrieved}}$$

Since both high recall and high precision are desired, the program was used to calculate the sum of these measures as well as the separate values, so that an over-all performance index might be obtained. To summarize the results, averages of these measures were then calculated over the whole set of document requests for each of the SMART processing methods, as well as over subsets of general and specific requests. (A general request is, roughly, one in which the number of relevant documents is at least 10.)

4. Data Sets Examined

The data used in the computations consisted of the ranked correlation coefficients obtained by eight of the SMART processing methods, four of these methods for a total of 24 requests, and the remaining four methods for 17 requests. The basic document file against which the requests had been compared consisted of a set of 405 abstracts of documents published during 1959 in the IRE Transactions on Electronic Computers. Of the requests eight were judged to be of a general nature and the rest specific. (This is the particular grouping of the requests used by J. Rocchio and M. Engel in their report on retrieval results obtained with the SMART system.[*])

In order to reduce the amount of keypunching necessary, it was assumed that the number of correlation values examined for each case could be taken as only slightly larger than the maximum number of relevant documents associated with any request. In particular, it was assumed that the 35 highest ranking correlation coefficients would be sufficient, and that the inclusion of additional points would not change the results. This assumption will be found valid by inspecting the data.

Each set of 35 points was processed by the following ten combinations of cut-off methods (listed in Part 3) and parameters, denoted as runs "A" through "J":

[*] Rocchio, J., and Engel, M., "Test Design and Detailed Retrieval Results" Information Storage and Retrieval, Report No. ISR-8 to the National Science Foundation, Computation Laboratory of Harvard University (December 1964).

Runs	Method	Parameter
A	1	.35
B	2	.02
C	2	.015
D	3	
E	4	
F	5	.02
G	5	.015
H	6	.05
I	6	.025
J	6	.02

For each cut-off method and for each of the eight SMART processing methods and a basic set of 17 requests, the recall, precision, and over-all measures were averaged over;

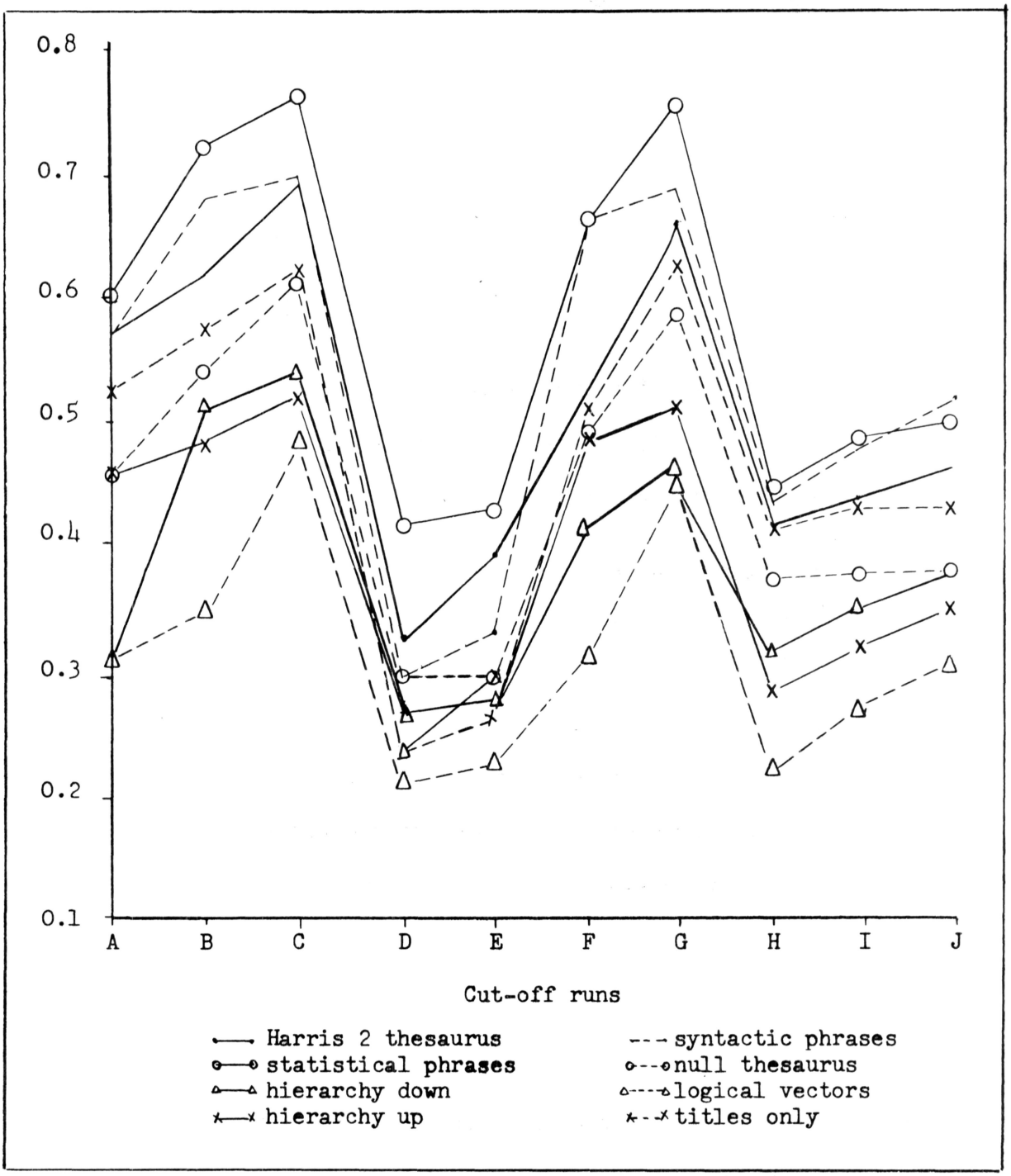
(a) all requests
(b) all specific requests
(c) all general requests

These results were then averaged over the eight SMART methods. For the four SMART methods for which seven additional specific requests were available, separate averages were calculated over a total of 16 specific requests to give some indication of the statistical validity of the results.

5. Results and Conclusions

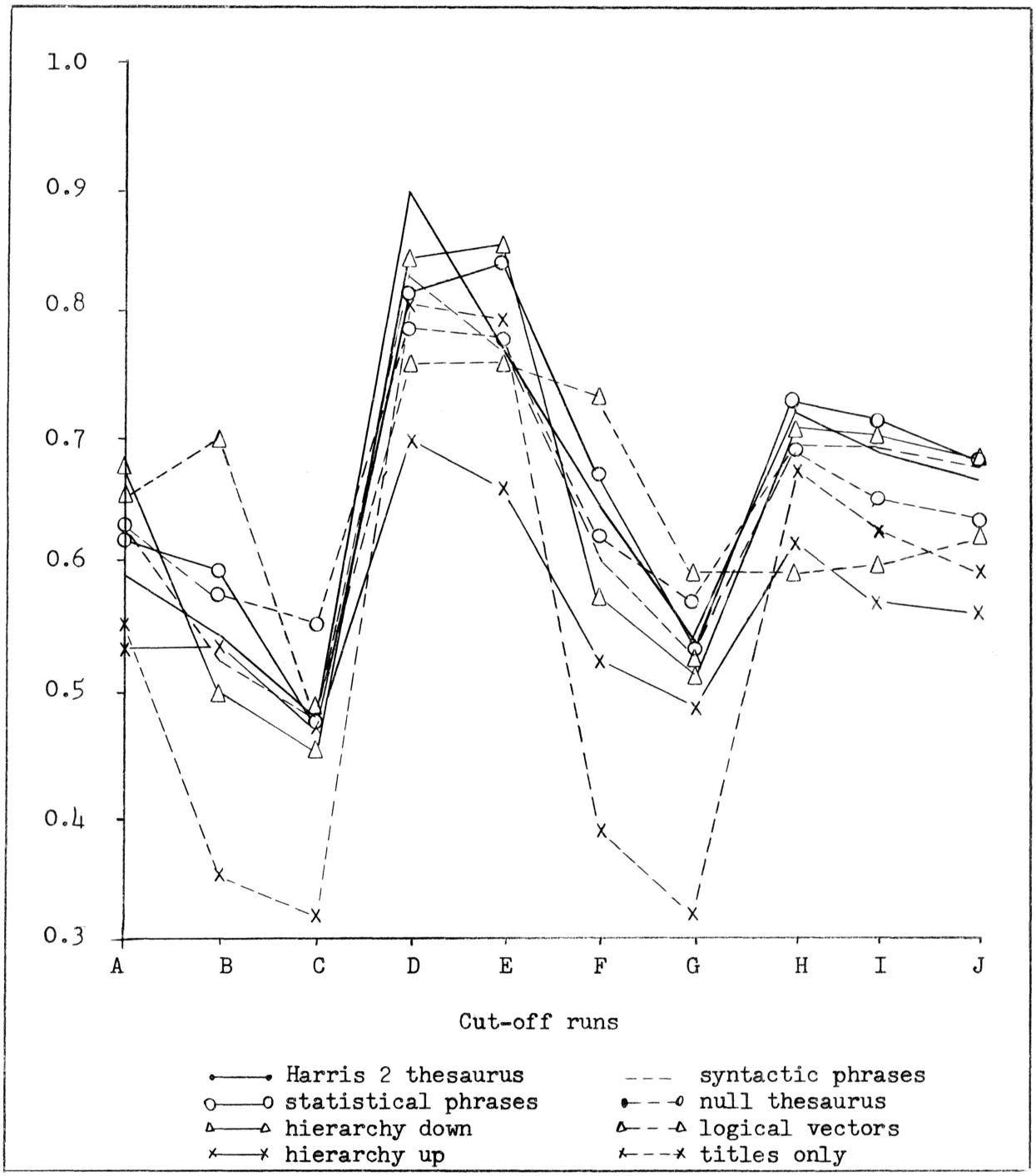
The results of these averaging procedures are shown in Figs. 2-6 from which the following conclusions may be drawn:

1. The relative performance of the cut-off methods does not depend heavily upon the underlying SMART processing methods since the curves in Figs. 2 and 3 cross infrequently, and retain a consistent shape.



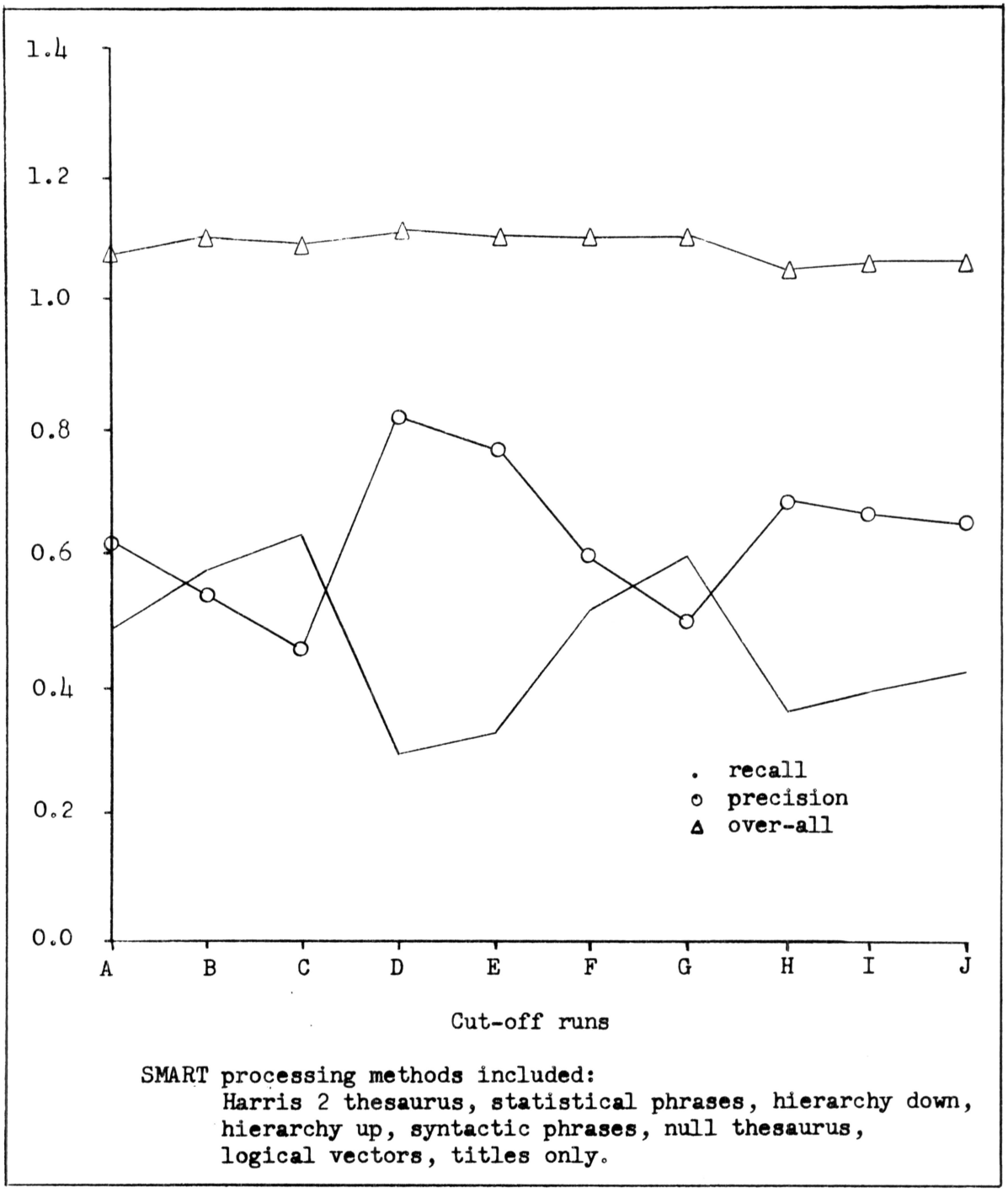
Recall Averaged over 17 Requests

Figure 2



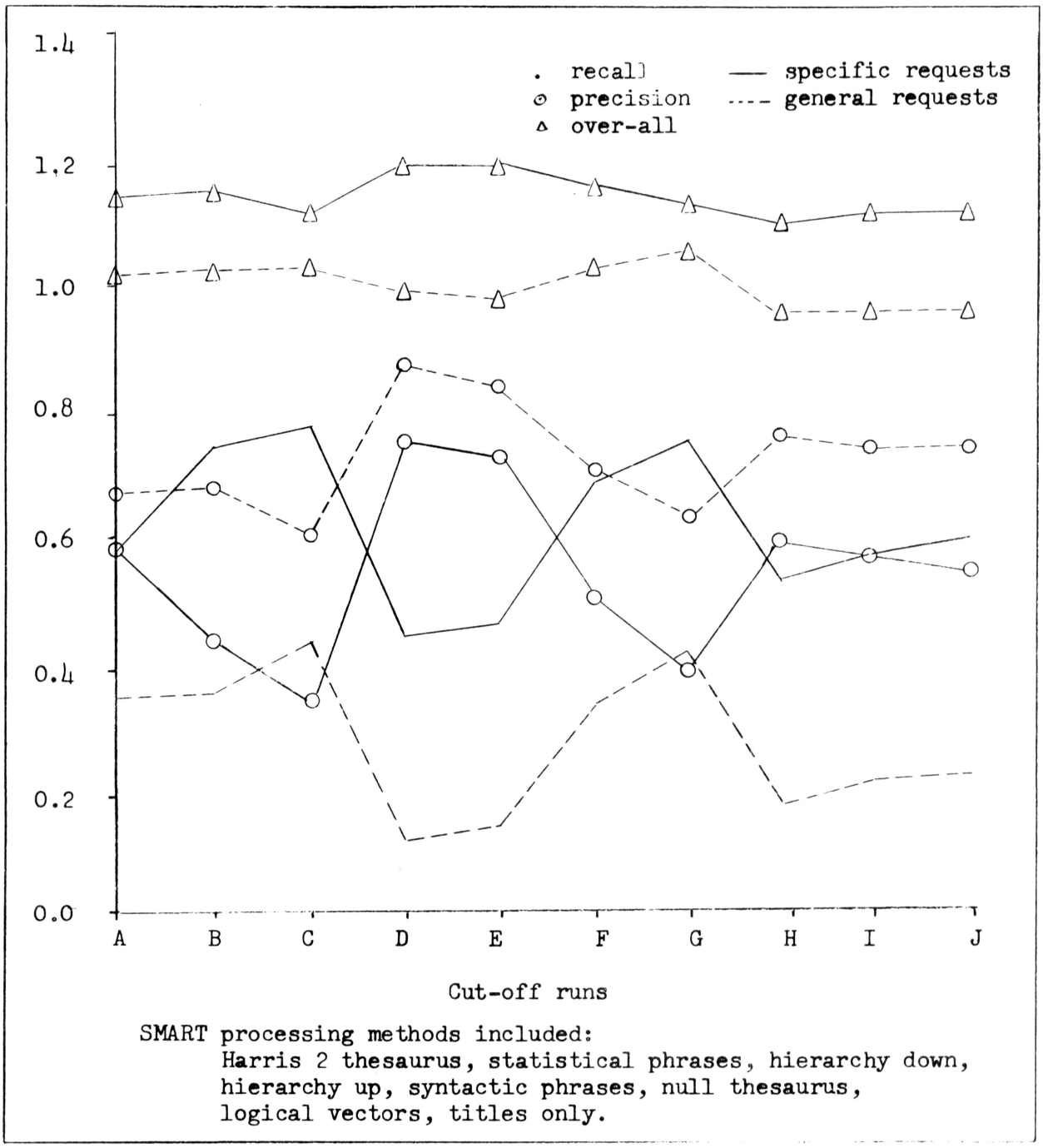
Precision Averaged Over 17 Requests

Figure 3



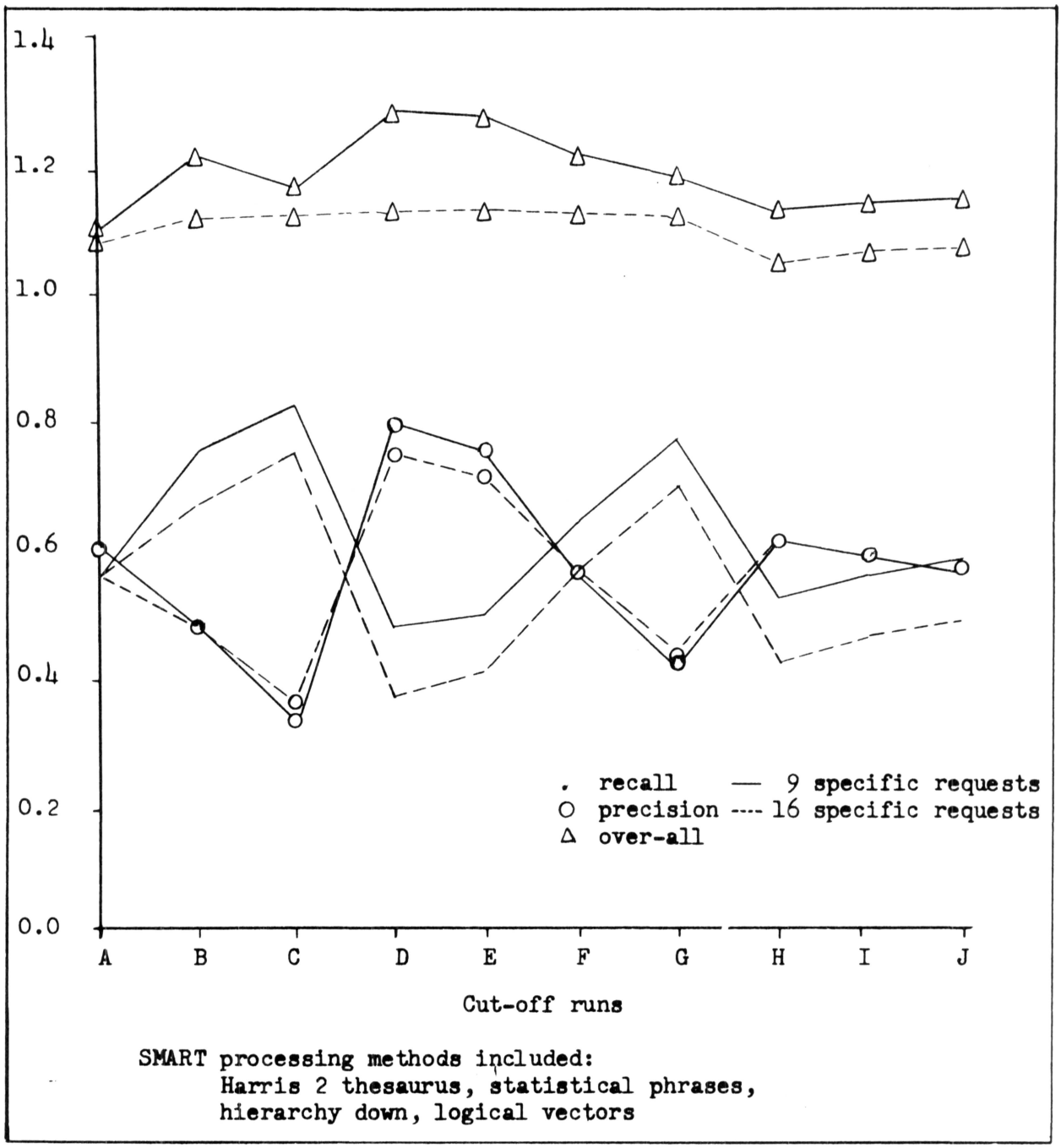
Averages Over Eight SMART Processing Methods
and 17 Requests

Figure 4



Averages Over Eight SMART Methods for Nine Specific and Eight General Requests

Figure 5



Averages Over Four SMART Methods, Specific Requests

Figure 6

2. There is a definite inverse relationship between recall and precision, low recall corresponding to high precision, and conversely.

3. This inverse relationship accounts for the surprising flatness of the curves for the over-all statistic.

4. The same cut-off methods produce the best precision and recall values for both specific and general requests, although the relative heights of the curves leads to an inverse relationship for the over-all statistic.

5. Higher recall values account for the fact that the over-all statistic is higher in the case of specific requests than for general requests. (The precision is actually higher for the general requests.)

6. Increasing the number of specific requests did not change the results substantively, although the over-all statistic was smoothed.

7. Method 6 which tends to even-off the number of documents retrieved from case to case is the least satisfactory of the methods.

8. None of the more complicated methods produced over-all results substantially superior to method 1 which is perhaps the most intuitively pleasing of the methods.