V. Evaluation Of Retrieval Performance

A. Measures Of Performance

Several average measures of the performance of the tested retrieval algorithms on the 42 Cranfield queries are used in this report. Each measure is based on the concept of "recall" and "precision". In evaluating an information retrieval system, an arbitrary cut-off point, such as rank ten or cosine correlation 0.75, is often employed. Documents above this cut-off point in the ranked list resulting from a search operation are considered "retrieved". With such a cut-off, recall is the percentage of documents relevant to the user that are retrieved, and precision is the percentage of retrieved documents that are relevant.

An ideal retrieval system would provide recall and precision of 100%, indicating that all relevant documents are retrieved and no non-relevant documents are retrieved. In SMART experiments an inverse relationship between recall and precision is observed, such that high recall implies low precision and vice-versa.

The 'document curves' used in this report are graphs of recall and precision at several cut-off points based on rank; that is, recall and precision after x documents are retrieved, for several values of x. The other measures used are not based on specific cut-off points, but in a sense measure retrieval performance over the entire document collection.

Normalized recall and normalized precision are two measures proposed by Rocchio [9] that take the average recall and precision obtained for all possible cut-off points. If N is the number of documents in the collection, R_j is the recall at a cut-off of j documents (rank j) and P_j is the precision at a cut-off of j documents, normalized recall and precision are defined as follows [15],

$$NR = \frac{1}{N} \sum_{j=1}^{N} R_j$$

$$NP = \frac{1}{N} \sum_{j=1}^{N} P_j$$

For automatic calculation, the following approximations are used in the SMART system [15]:

$$NR = 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n\,(N-n)}$$

$$NP = 1 - \frac{\sum_{i=1}^{n} \ln r_i - \sum_{i=1}^{n} \ln i}{\ln \left(\frac{N!}{n!\ (N-n)!} \right)}$$

where r_i is the rank of the i^{th} relevant document in the collection and n is the number of relevant documents in the collection for the given query. A normal overall measure

of retrieval performance has been suggested [15] but is not explicitly displayed in this report: Normal overall measure = 1 - 5 NR + NP. The factor of 5 gives equal weight to the two component measures.

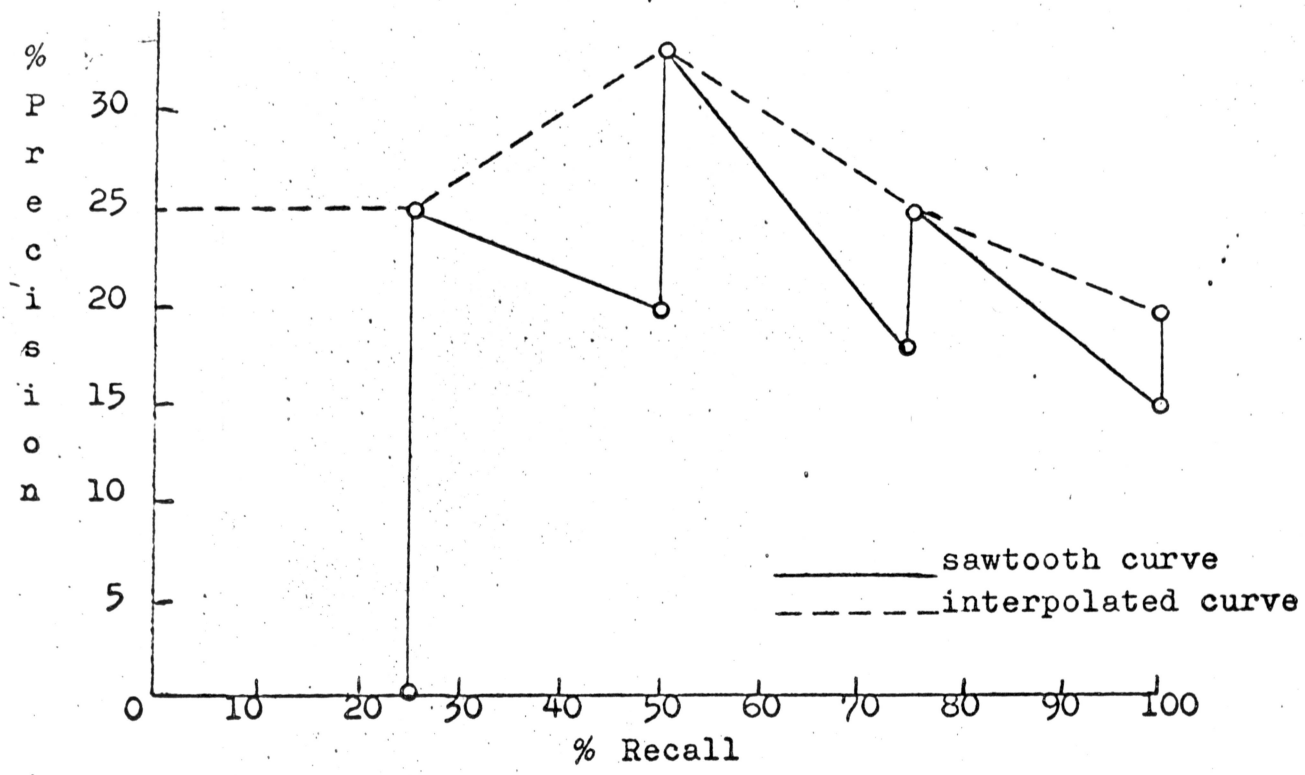
Another overall measure used in many studies of retrieval performance is the recall-precision curve, an average plot of precision at each 5% or 10% of recall. Each query is averaged into each point of the plot. To accomplish this averaging process, an interpolation procedure is needed, since, for example, a query with two relevant documents can only achieve uninterpolated recall levels of 50% and 100%.

Two types of recall-precision curve are used in this study. They are distinguished by the method of interpolation used. Both the Quasi-Cleverdon interpolation used in several previous studies and the Neo-Cleverdon interpolation now used for all evaluation of the SMART system are described below.

Figures 1 and 2 show two graphs for a hypothetical query having 4 relevant documents. The relevant documents are assumed to be retrieved with ranks of 4, 6, 12 and 20. Thus, at 25% recall, the precision is 25%, at 50% recall, the precision is 33%, and so on. However, these values correspond actually to the highest possible precision points, since they are calculated just after a relevant document is retrieved. In this example, after 3 documents are retrieved,

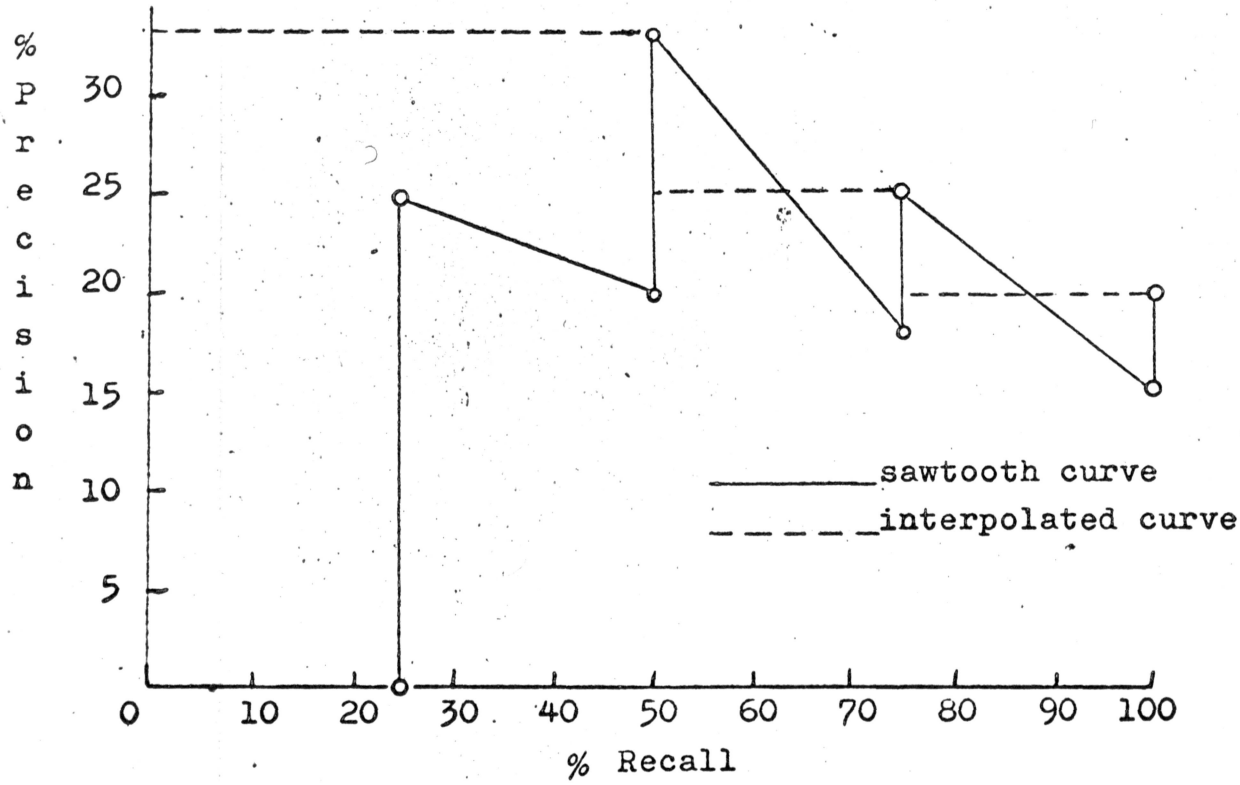
the precision is 0%, after 5 documents, the precision is 20%, and so on. This range of precision for each recall level is indicated by the top and bottom points in Figures 1 and 2 at 25%, 50%, 75%, and 100% recall. The solid saw-tooth line connecting these points is not used for interpolation; it is intended to indicate the drop in precision between the actual recall levels for this query as more non-relevant documents are retrieved.

The Quasi-Cleverdon interpolation uses a straight line between peak points of precision, as indicated by the dashed line in Figure 1. It has been argued that this interpolation is artificially high, since it lies at all points above the saw-tooth curve, and thus, does not reflect in any way the precision drop as more non-relevant documents are retrieved. The Neo-Cleverdon interpolation of Figure 2 projects a horizontal line leftward from each peak point of precision, and stops when a higher point of precision is encountered. This new interpolation curve (the dashed line in Figure 2) does <u>not</u> lie above the saw-tooth curve at all points. When the precision drops from one recall level actually achieved to the next, an immediate drop in precision after the first point to the level of the next point is indicated. For example, in Figure 2, the precision value at 50% recall is 33%, but at 55% recall, the interpolated value used for the new averages is 25% precision. When the precision rises from one recall level to the next, however, the first precision point actually achieved is ignored for



An Illustration of the Interpolation Method Used for the Quasi-Cleverdon Recall-Precision Averages

Figure 1



An Illustration of the Interpolation Method Used for the Neo-Cleverdon Recall-Precision Averages

Figure 2

purposes of interpolation. The achieved precision of 25% at 25% recall in the example of Figure 2 is ignored, and for all recall levels from 0 to 50%, an interpolated precision of 33% is used for the new averages. The proponents of the new interpolation argue that this method indicates in all cases a precision that the user could actually achieve, if he were to use clairvoyance to retrieve exactly the right number of documents.

B. Statistical Significance Tests

Several statistical tests are reported here using as input the rank recall, log precision, normalized recall, normalized precision, and 10 points from the feedback effect (Section V-C) recall-precision curve with the Neo-Cleverdon interpolation. The statistical tests are intended to measure the "significance" of the average difference in values of these measures obtained for two iterations or two distinct search algorithms. The test results are expressed as the probability that the two sets of values obtained from two separate runs are actually drawn from samples which have the same characteristics. A small probability value thus indicates that the two curves are significantly different. If this probability for one measure is, for example, 5%, the difference in the two average values of that measure is said to be "significant at the 5% level".

Choice of a statistical method for calculating this probability is important. The present study uses three statistical tests, the familiar T-test, the Wilcoxon Signed-

Rank Test (WSR), and the Wilcoxon Rank Sum Test (WRS) [16].

The T-Test and the Wilcoxon Signed-Rank Test are used in this report to compare the retrieval of one feedback iteration to another or of one algorithm to another, using all queries. The T-Test takes account of the magnitude of the differences, and assumes that the measures tested are normally distributed. The WSR test does not make this assumption. Moreover, the WSR test takes account only of the ranks of the differences, ignoring their magnitude. Because this test does not assume normality of the input and because it ignores some information (magnitudes of differences), the WSR test is more conservative than the T-Test. It is therefore less prone to the error of calling a result "significant" when it is not. Because information retrieval provides discrete rather than continuous data, and because only 42 data points (42 queries) are provided, the more conservative WSR test is preferable for the present evaluation.

The Wilcoxon Rank Sum Test can be used to test unpaired observations, and is used in Section VI-E of this study to compare one subgroup chosen from the 42 queries to a contrasting subgroup of queries. Like the WSR test, the WRS test ignores the magnitudes of the results and does not assume a normal distribution.

C. The Feedback Effect in Evaluation

The assignment of ranks to documents retrieved for feedback is a key factor in the evaluation of retrieval performance.

Two methods of assigning these ranks have been proposed, and both are used in the present study. Hall and Weiderman [17] compare and evaluate these two methods. In previous feedback investigations, all documents in the collection received new ranks after each iteration and the top-ranked N documents were used for feedback. Hall and Weiderman point out that evaluation of this retrieval technique takes into account two effects, which they call "ranking effect" and "feedback effect".

Relevance feedback in effect uses information from one or more document descriptors to modify the query descriptor. The relevant documents used for this purpose will be ranked higher by the modified query than previously, and the non-relevant documents used will be ranked lower. The effect of these rank changes in "retrieved" documents is termed the "ranking effect". If the ranking effect is included in an overall performance measure, the measured change in performance between feedback iterations is quite impressive.

This large change in "total performance" (including both ranking and feedback effect) indicates the extent to which the initial query has been perturbed toward the centroid of the relevant documents, and strongly supports Rocchio's theory.

Hall and Weiderman state that in an environment where the user must actively supply relevance judgments for feedback, changes in the ranks of documents which the user has already seen are of no interest to him. The user in such an

environment is concerned primarily with the "feedback effect"; that is, the effectiveness of the modified query in bringing new relevant documents to his attention. They conclude that, though total performance is a valid measure of the effectiveness of relevance feedback in approaching the "ideal query", the feedback effect should be isolated and examined as well.

The present study evaluates total performance and also measures feedback performance in the manner suggested by Hall and Weiderman, discarding the ranking effect and presenting only the feedback effect. The ranks of the top N documents retrieved in each iteration (the documents used for feedback) are "frozen" in all subsequent iterations, and only the remainder of the collection is searched using the modified query. Thus, in feedback effect evaluation, the N documents retrieved on any iteration are guaranteed to be N <u>new</u> documents; that is, documents <u>not</u> used for feedback on any previous iteration. Moreover, the performance measures for the first (second, third) iteration are calculated from a ranked document list in which the top N (2N, 3N) documents are the same as those retrieved previously. Only the changes in the ranks of documents not yet seen by the user is measured.

Feedback effect evaluation gives overall results that are deceptively low. Because the top ranks are frozen, no newly retrieved document can achieve a rank higher than that of any previously retrieved document. With a constant

feedback strategy, therefore, on the first (second, third) iteration, the highest possible rank for a new document is N+1 (2N+1, 3N+1). For this reason, the feedback effect evaluation is a misleading measure of the overall performance of the retrieval system, and should be used in conjunction with other evaluation methods. Islation of the feedback effect is primarily useful to compare different feedback strategies from the viewpoint of a user in an interactive retrieval environment. Figure 35 in Section VII-B compares total performance and feedback effect evaluation of similar feedback algorithms.

However, one feature of feedback effect evaluation is psychologically essential to a realistic relevance feedback system; the guarantee that the N documents retrieved on any iteration have not previously been seen by the user. For this reason, new evaluation methods that provide this guarantee without severely limiting the attainable retrieval performance should be investigated. Several such methods are discussed in Section VII-B.

The results reported in this study include:

Total Performance:

1. Normalized recall and precision
2. Recall-precision curves with Quasi-Cleverdon interpolation.

Feedback Effect:

1. Normalized recall and precision
2. Recall-precision curves with Neo-Cleverdon interpolation.
3. Document curves at several cut-off points

4. T-tests and Wilcoxon Signed Rank tests of the normalized measures and of recall-precision curves.
5. Wilcoxon Rank Sum tests of normalized recall and precision.