ISR-9
August 1965

XIX. COMBINATIONS OF ANALYSIS METHODS - THE MERGED OUTPUT RESULTS

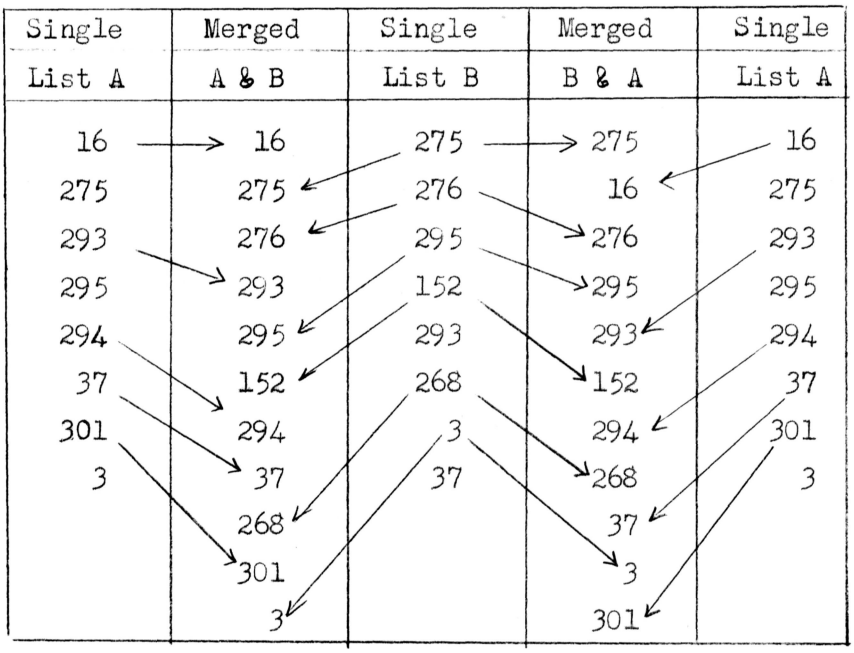
J. J. Rocchio, Jr.

1. Introduction

One of the most interesting features of automatic indexing in document retrieval is the flexibility it permits in the generation of index images of documents and search requests. The SMART system is designed to take advantage of this flexibility, and one of its principal aims is to study the effect of various information analysis procedures on retrieval performance. In addition to investigating the gross retrieval characteristics of various indexing techniques, it is interesting to examine their fine structure as reflected, for example, in their retrieval performance with respect to particular relevant documents. This is particularly appropriate in light of the fact that an operational retrieval system may require a variety of techniques adaptable to a wide spectrum of users' information needs. One of the means for measuring the detailed behavior of a particular sequence of analysis methods is to consider their joint behavior under the assumption that a given user will first use method A, and will then switch to method B if the original search results were found wanting. This method has been investigated by using some of the initial results obtained with the SMART system after "merging" the retrieval results obtained from different procedures into an effective combined result. The present section outlines the merging technique, and describes some of the results which have been obtained.

2. The Merging Process

The details of the merging procedure for the retrieval results obtained from various analysis methods, as well as characteristics of the computer program which performs the merging operation, are described in Sec. XVIII of this report. In summary, the retrieval results for an experimental run using the SMART system consist of an ordering of the documents of the source collection with respect to the input query. The combined result for m-different methods is defined by an m-way merge of the ordered list of document identifiers pertaining to each individual method, such that in the merged list each document identifier is represented only once with a rank determined by its highest rank in one of the original lists (see Fig. 1).



Example Illustrating the Two Possible Merges of Two Individual Lists

Figure 1

The result of such an m-way merge is then again an ordered list of document identifiers, which is taken to be the composit or effective result of the m methods combined. Such a merge is somewhat sequence dependent, in that the resultant ordering is a function of the sequence in which the m component methods are considered (see Fig. 1). However, when m is small compared to the number of items to be merged, this sequence dependency does not seriously affect the evaluation measures derived from the resultant list.

By considering the nature of this merging process, some general conclusions can be drawn. In particular, a document which has a minimal rank index i in a set of m-individual methods can have at most a rank index of m - i in the ordered merged list. Thus the rank position of a document in a combined method is a function of its best ranking (lowest rank index) over all component methods. This is illustrated in Fig. 2, which shows the ranks of the relevant documents for the request "Automata Phrases," for each of three simple methods, and for the merged result of all three. It should be noted, for example, that document 264 is ranked 74th under "Harris TWO," 64th under Stat. "Phrases," and first under "Hierarchy Up." In the combined result it receives rank 3, indicating that "Hierarchy Up" was last in the merging order. It is clear that the rank list of relevant documents will be improved by the merging process, if the individual methods are successful in identifying <u>different</u> subsets of the relevant set for the input query. In the case where a given method dominates all others, that is, retrieves every relevant document with the highest rank (lowest rank index), the merging procedure is likely to produce worse, or certainly not better combined

Query: Automata Phrases

Relevant Document Number	Rank Position in Retrieved Ordering			
	Harris Two	Hierarchy Up	Stat Phrase	H2 + Up + Stat
316	1	2	2	1
129	2	5	9	4
313	3	22	5	5
176	4	50	8	8
371	7	68	1	2
372	29	116	3	6
241	38	133	7	15
315	42	196	39	74
264	74	1	64	3

Ranks of Relevant Documents for Search Request "Automata Phrases" Under Three Analysis Methods and the Merged Result

Figure 2

results. Comparisons of the combined retrieval effectiveness for sets of alternative analysis procedures provide a measure then of the potential usefulness of incorporating these methods into an operational retrieval system.

3. Sample Results

Figure 7 of Sec. XXIV of this report presents in summary form results of SMART system retrieval experiments for a sample of 17 search requests. Data are presented for six analysis procedures, including the thesaurus mapping (Harris Two), the word stem mapping (Null Thes.), the addition of phrases

detected by concept-concept cooccurrence (Stat. Phrases), the addition of phrases detected by syntactic matching (Syntax Phrases), the search request transformation by means of adding related "more general" terms from the concept hierarchy (Hierarchy UP), and the request transformation by means of adding related "more specific" terms from the concept hierarchy (Hier. down). In addition, 13 two-way merges and nine three-way merges of these six basic methods are included. Note that all of these indexing methods are transformations of the concept space defined by the thesaurus with the exception of the method labeled "Null Thesaurus." This last indexing transformation is for a distinct space defined by the detection of unique word stems from the source collection.

The summary data for the single and combined methods may be presented in the form of "Quasi-Cleverdon" graphs, or as averages of various evaluation measures (see Secs. III and IV of Information Storage and Retrieval, Report No. ISR-8). Figure 3 presents a summary of the relations of the various merges to their component measures. In this figure improvement, represented by a "+," indicates that a higher normalized recall or precision value was obtained for the merged method compared to the individual simple ones. It is evident that with the exception of some of the merges involving the two hierarchy methods, the merging process provides general improvement.

In terms of the normalized precision evaluation measure (the more sensitive of the two normalized measures), the best single method is the "Stat. Phrase" option. The best merged pair is obtained by combining the "Stat. Phrase" method with the "Null Thes." method, and this in fact turns

	Null		Stat		Syntax		Hier Up		Hier Down	
Harris Two	-	+	+	+	+	+	-	-	-	-
Null			-	+	+	+				
H2 + Null							-	-	+	+
H2 + Stat	-	+					-	-	-	-
H2 + Syntax	+	+					-	-	-	-
H2 + Up	+	+	-	-	+	+			+	+
H2 + Down	+	+	-	+	+	+	-	-		

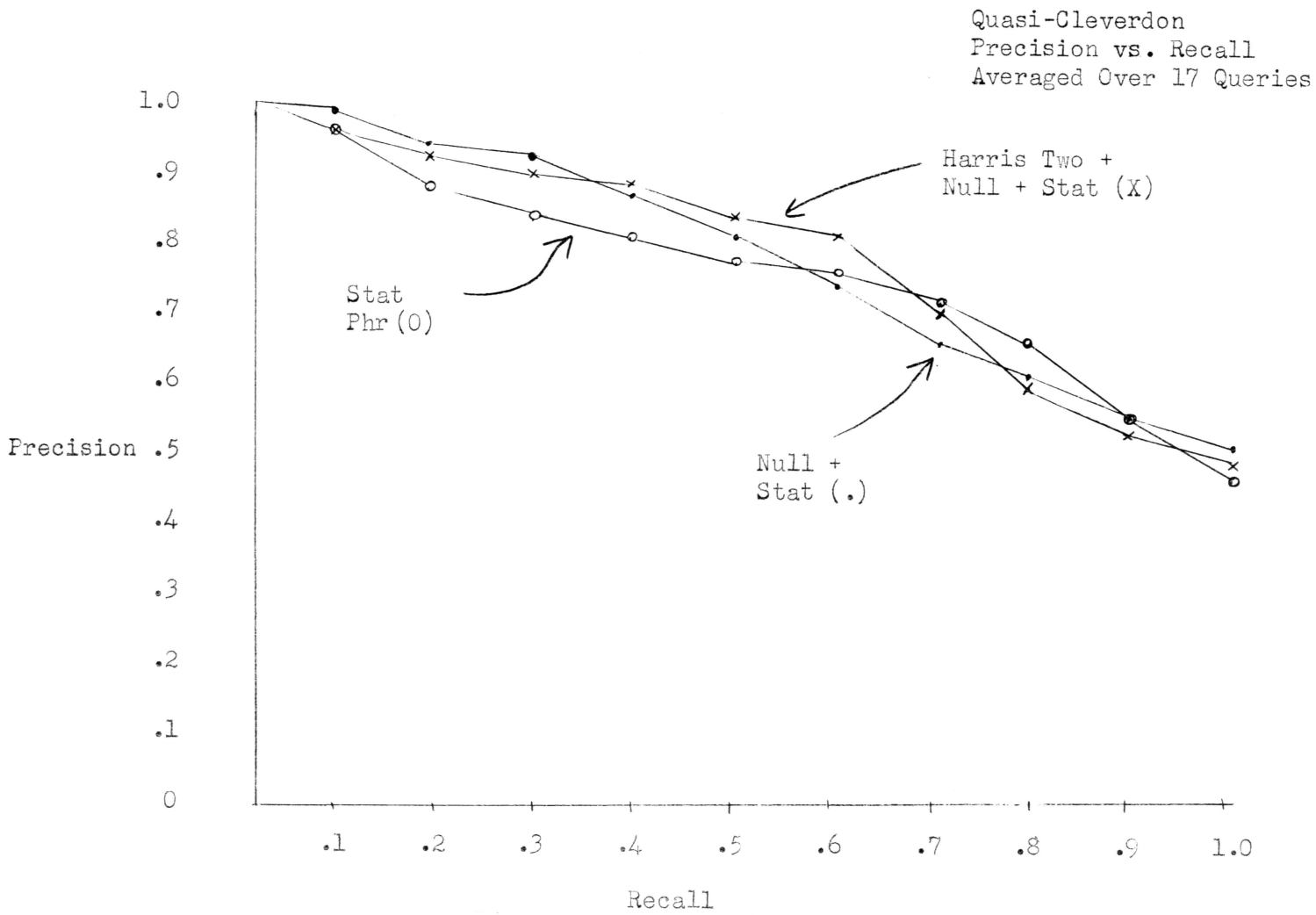
(a) Increases (+) or decreases (-) in normalized recall (left) and normalized precision (right) for certain merged methods

	Null + Stat		Null + Syntax		Stat + Hier Up		Stat + Hier Down		Syntax + Hier Up		Syntax + Hier Down		Hier Up + Hier Down	
Harris Two	+	-	+	+	+	+	-	*	-	+	-	+	-	-

(b) Increases (+) or decreases (-) in normalized recall (left) and normalized precision (right) for Harris II and certain merged pairs (* indicates no charge).

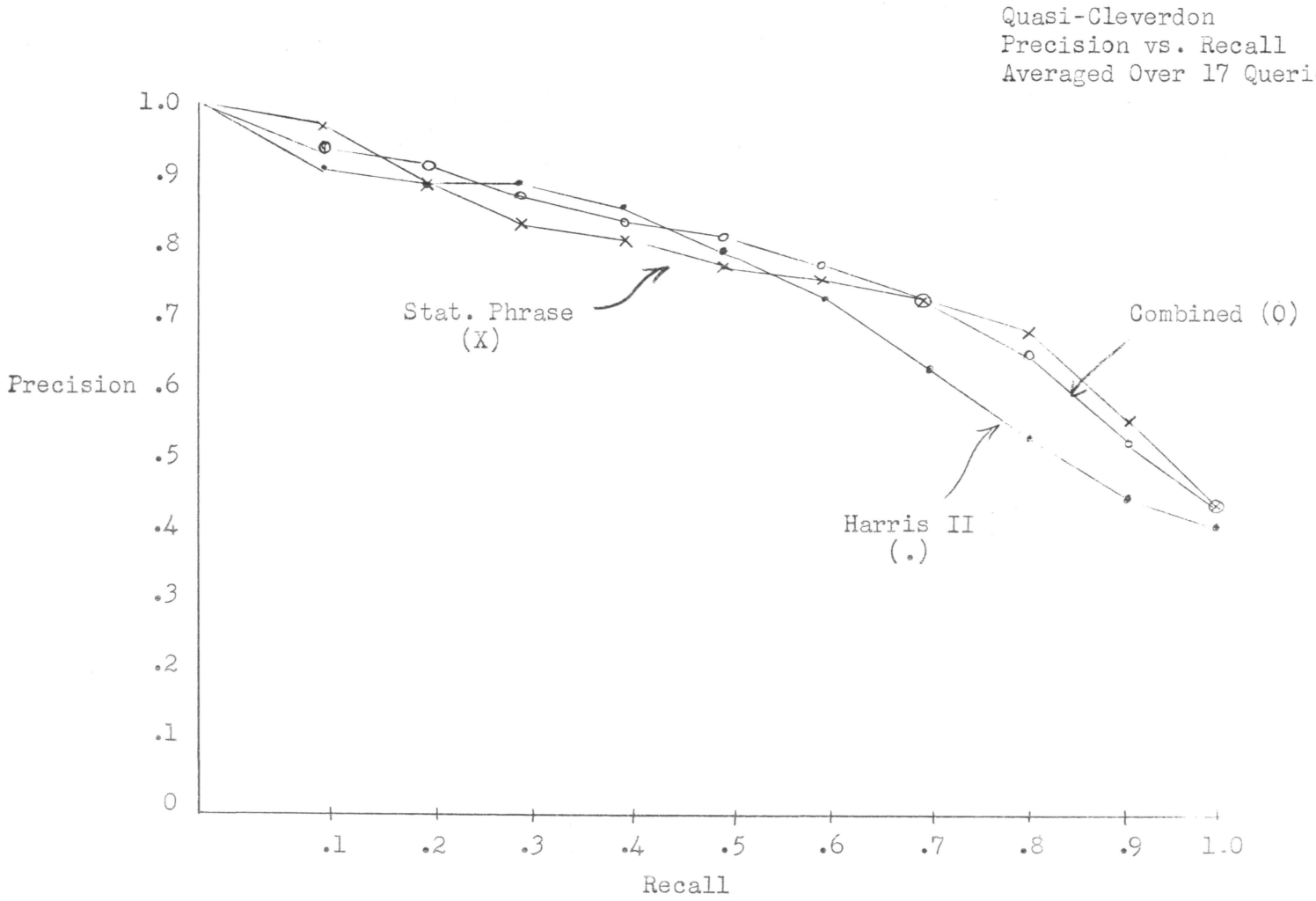
Results of Various Merges

Figure 3



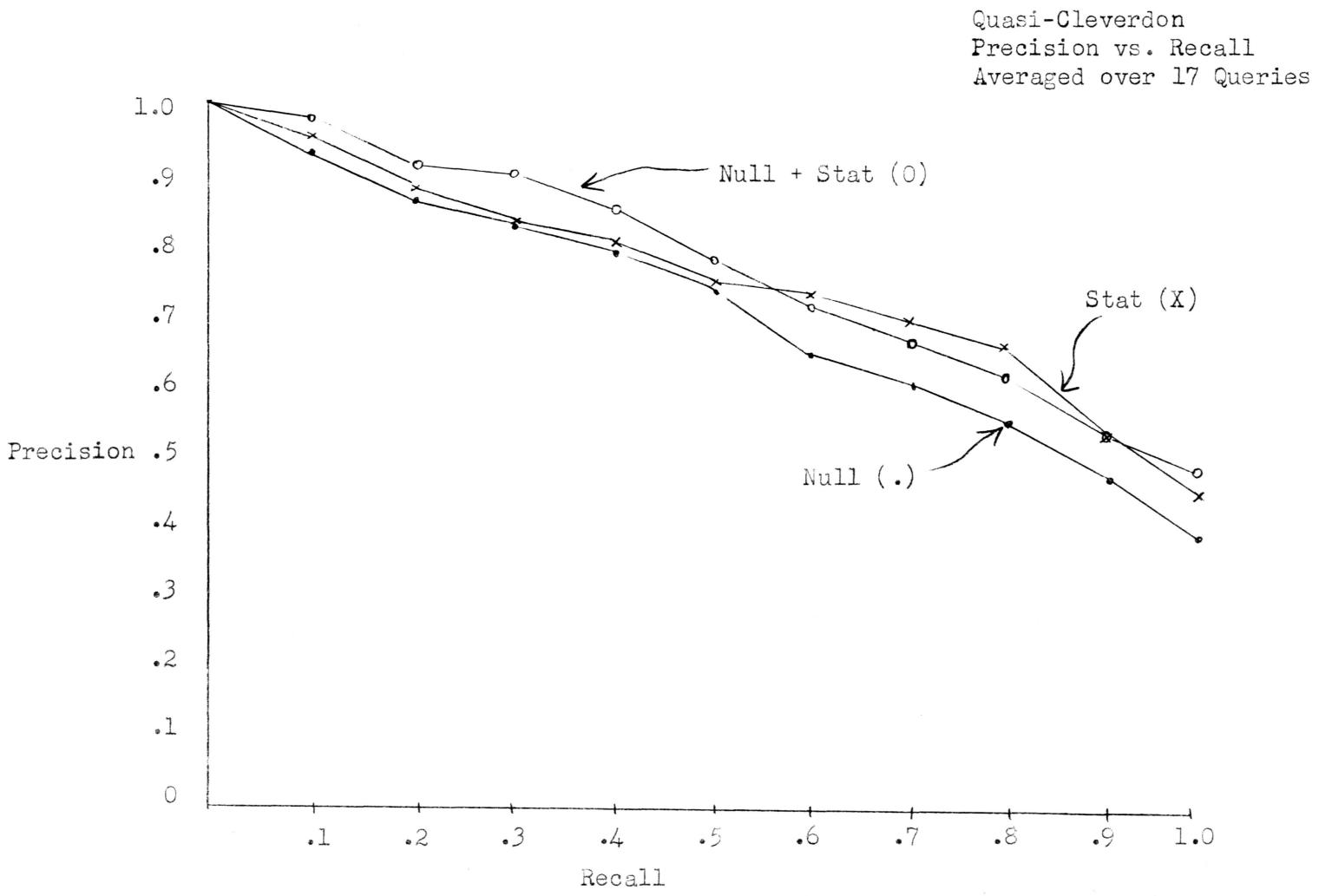
Comparison of the Best Single-, Double-, and Triple-merged Methods

Figure 4



Quasi-Cleverdon Precision Versus Recall for Harris Two,
Stat Phrase, and Their Combination

Figure 5



Quasi-Cleverdon Precision Versus Recall for Null Thes, Stat Phrase, and Their Combination

Figure 6

out to give the maximum normalized precision over all 27 methods presented. The best three-way merge results from the combination of the "Harris Two" method with the "Null Thes." and "Stat. Phrases" methods, and this yields the second highest average normalized precision. The "Quasi-Cleverdon" precision versus recall graphs for these three methods are shown in Fig. 4. Figures 5 and 6 show the precision versus recall curves for Harris Two, Stat. Phrases and their merge, and Null Thes., Stat. Phrase and their merge, respectively. In each of these cases it is important to note that at certain recall levels the merged method results in higher precision than either of its components, the improvement being more marked in Fig. 6.

4. Conclusions

The results obtained to data, illustrated in the present section, support the hypothesis that combinations of alternative information analysis procedures can provide improved retrieval characteristics. Specifically, the merging of certain of the retrieval options available in the SMART system have shown marked improvement over the individual component methods. These results suggest that document retrieval systems be equipped with a set of analysis procedures capable of being employed in user directed sequences, as opposed to using a single, fixed indexing technique.