IV. Environment Of The Reported Experiments

The present study of relevance feedback compares several related formulas for query modification. The following general formula is available to the experimenter:

$$Q_{i+1} = \pi Q_i + \omega Q_o + \alpha \sum^{\min(n_a, n_r)} r_i + \mu \sum^{\min(n_b, n_s)} s_i \quad (D)$$

where $n_r + n_s$ (see formula A, Section 1) equals N, the number of documents retrieved for feedback.

The experimental variables are α, ω, π, μ, n_a, n_b, and N. The parameter α is positive, and weights all incoming relevant documents relative to the other contributors to the query (previous query, initial query, non-relevant documents). The parameter π permits the previous query to be increased in weight relative to the incoming documents. Q_o is the initial query, as opposed to the query of the previous iteration; ω permits the initial query to be used as part of the new query (see Section 3B). The parameter μ should theoretically be negative, as it permits some significance to be attached to the non-relevant documents retrieved. The parameter n_a (n_b) permits some specific number of relevant (non-relevant) documents to be used in the query even if n_r (n_s) is larger. It is assumed that the r_i and s_i are indexed in order of decreasing relevance (as determined by the system) to the query; that is, the n_a relevant documents (or n_b non-relevant documents) used in the new query will be those

closest in the descriptor space to the previous query. The flexibility of this formula permits the investigation of several feedback strategies.

The system also provides the following formula to simulate Rocchio's algorithm:

$$Q_{i+1} = \pi\, n_r n_s Q_i + \omega\, Q_o + n_s \sum_{r_i - n_r}^{\min(n_r,\, n_a)} - \sum^{\min(n_s,\, n_b)} s_i \quad (E)$$

Formula E does not normalize the vector lengths as is done in Rocchio's algorithm (formula A).

The document collection used in this study (the "Cranfield" collection) contains 200 documents from the field of aerodynamics, chosen from a library of 1400 documents. For this collection, there are 42 queries, constructed by some of the authors of the 1400 documents; these requestors are also responsible for the relevance judgments.

The concept vectors describing document and queries are quite sparse for the "Cranfield" collection. The maximum number of concepts used to describe one document is 85, out of a possible 552 concepts. The largest weight given to any concept in any document descriptor is 288. The query description vectors are sparser by one order of magnitude and shorter than the document descriptors. The maximum number of concepts used in a single query vector is 13; the largest weight in any query vector is 24. The largest number of documents relevant to a single query is 12, or six percent of the collection. The comparative brevity of

the query vectors in this collection is typical in technical document retrieval, because document abstracts generally contain more detailed information than user queries.

The characteristics of an experimental document collection determine the extent to which experimental results typify retrieval behavior in 'real' document collections. The three collections described in this study are much too small to require sophisticated retrieval techniques. However, the Cranfield 200 document collection is more realistic than the ADI and IRE collections for three reasons.

First, more queries are available for the Cranfield 200 collection. The number of queries available is important in judging the statistical significance of experimental results.

Second, the documents in the Cranfield 200 collection were chosen from a more typical environment. The ADI collection consists of short papers all presented at the same conference. The papers in the IRE collection were all published in the same magazine within a seven-month period. By contrast, the 1400 documents in the full Cranfield collection were in effect selected by knowledgable authors from the field of aerodynamics, and the 200 documents in the small collection were chosen to represent the larger collection.

Third, the queries and relevance judgments in the ADI and IRE collections were constructed by a small number of information retrieval experts, while those in the full Cranfield collection were constructed by 182 authors of

recent papers in aerodynamics.

Concept vectors for both the ADI and Cranfield collections were constructed automatically using a regular subject-area thesaurus. In all experiments reported here, the cosine correlation (Section I) is used as the distance function.