An Investigation of Relevance Feedback Using Adaptive Linear and Probabilistic Models

Robert G. Sumner, Jr., and W. M. Shaw, Jr.
School of Information and Library Science
University of North Carolina
Chapel Hill, NC 27599-3360 USA
{sumnr, shaw}@ils.unc.edu

Abstract

The SMART system (v. 11.0) was used as a front-end to a two-stage retrieval process. In the first stage, (WSJ) documents and the description field of (ad hoc) topics were indexed by the stems of single terms; *lnc* and *ltc* weights were computed for word stems in documents and queries, respectively; and documents were ranked according to the cosine similarity of document and query vectors. Related by the initial query vector, the first 5000 documents in the ranked list for each topic constituted a "condensed" database for that topic. Preliminary experiments with TREC-4 topics and official relevance evaluations suggested each such database would include a high fraction of relevant documents for the associated topic, and the result was confirmed by TREC-5 results. In the second stage, initial query vectors were automatically refined by two relevance feedback strategies applied to the condensed databases. One of us employed the adaptive linear model (*uncis1*), and the other used a variation of the "classic" probabilistic model (*uncis2*); relevance judgments were made independently. In *uncis1*, the query at a given search iteration is expanded by all terms in relevant, retrieved documents and all terms in selected, nonrelevant, retrieved documents, and documents are ranked by the inner product of document and query vectors. In *uncis2*, the query is expanded by all terms in relevant, retrieved documents, and documents are ranked by the cosine similarity of document and query vectors. For *uncis1* and *uncis2*, respectively, average non-interpolated precision values over all relevant documents are 0.25 and 0.20, and average R-precision values are 0.25 and 0.21. Results show that independent relevance judgments made in *uncis1* and *uncis2* are quite different and have a strong effect on retrieval outcomes; our relevance evaluations also differ significantly from official relevance judgments. Retrieval performance improves when official relevance judgments are utilized by both models. For the 31 topics in which there was an official relevant document in the top 34 of the initial ranking, average non-interpolated precision values are 0.60 for the adaptive linear model and 0.59 for the probabilistic model.

Introduction

The next generation of information retrieval (IR) systems can be expected to be interactive, adaptive, and intelligent. Such systems will allow end-user relevance assessments and end-user query modifications to guide the construction of effective representations of information need. Incorporation of relevance feedback in IR experimentation has been found to improve retrieval performance as shown, for example, by Buckley, Singhal, Mitra, and Salton (1996); Haines and

Croft (1993); Robertson, Walker, Beaulieu, Gatford, and Payne (1996); and Salton and Buckley (1990). Among numerous unresolved issues related to end-user and IR system interactions, we addressed two questions in our experiments for TREC-5.

(1) Is the effectiveness of relevance feedback adversely affected by concentrating on a small subset of a document collection defined by an initial query statement?
(2) What is the relative effectiveness of alternative relevance feedback models in the context of end-user relevance evaluations and automatic query expansion?

Two-Stage Process

Information retrieval has been viewed as a two-stage process (Boyce, 1982; Shalini, 1993). In the context of the present investigation, an initial ranking of documents is produced for a topic and constitutes the outcome of the first stage. In the second stage, the top *n* documents of the initial ranking serve as a "condensed" collection for further retrieval processing. For example, Buckley, Salton, Allan, and Singhal (1995) selected the first 1750 documents from a "global" search of the TREC-3 database, reranked the 1750 documents according to "local" criteria, and found that the resulting ad hoc run (*CrnlLA*) compared favorably with a run applied to the complete database (*CrnlEA*). If the top *n* documents related by an initial query vector include a high fraction of the documents relevant to that query, applying relevance feedback to the condensed collection offers an efficient strategy for producing a satisfactory retrieval outcome. If the top *n* documents do not include a high fraction of relevant documents and the retrieval outcome is not satisfactory, the resulting query vector, refined by several iterations of the feedback process, can be resubmitted to the full database for further processing. In our TREC-5 runs, the second stage of the retrieval process focuses on the use of relevance feedback strategies. We are testing the effectiveness of relevance feedback algorithms applied to a condensed collection for each topic; we have not yet explored the utility of submitting the refined query vector to the entire database.

The effectiveness of our feedback runs is influenced by the proportion of relevant documents among the top *n* documents of the initial ranking; if a relevant document is not in the top *n*, then it cannot be retrieved. To determine an appropriate value for *n*, we conducted preliminary experiments using TREC-4 ad hoc topics and official relevance assessments for the portion of the Wall Street Journal (WSJ) collection used in category B runs. An initial ranking of documents was created for each topic using the same ranking procedure that was employed in our TREC-5 runs. The proportion of relevant documents that are in the top *n* documents (i.e., recall) was determined for several values of *n*. For a given value of *n*, Table 1 shows the number of topics with a recall value within the stated range. The 45 topics evaluated are those with at least one relevant document in the collection. As shown in Table 1, 100% of the relevant documents appear in the first 5000 documents for 30 of 45 topics, and 80% or more of the relevant documents appear in the first 5000 documents for 43 of the 45 topics.

Table 1: For a given value of *n*, the number of TREC-4 ad hoc topics with the value for recall within the stated range.

Recall (*Rec*)	Value for *n*				
	2000	3000	4000	**5000**	6000
$Rec = 1.000$	21	26	28	**30**	30
$0.900 \le Rec < 1.000$	8	7	7	**7**	7
$0.800 \le Rec < 0.900$	7	6	6	**6**	6
$0.700 \le Rec < 0.800$	4	4	2	**0**	0
$0.600 \le Rec < 0.700$	2	0	0	**0**	0
$0.500 \le Rec < 0.600$	2	1	1	**1**	1
$0.400 \le Rec < 0.500$	0	0	1	**1**	1
$0.300 \le Rec < 0.400$	1	1	0	**0**	0

Initial Ranking of Documents

SMART (v. 11.0) was used to index the WSJ collection and the fifty ad hoc topics. For each document, <HL>, <LP>, and <TEXT> fields were indexed. These three fields correspond to the headline, lead paragraph, and remaining paragraphs, respectively, of a WSJ article. Only the <desc> field of the topics was indexed. SMART indexed the documents and topics by removing stop words and stemming the remaining terms with the modified Lovins algorithm.

In SMART, each document i is represented by a document vector $\mathbf{d_i}$, and the query is represented by a query vector $\mathbf{q}$. Each component d_{ik} of $\mathbf{d_i}$ is the weight of term k in document i. Likewise, each component q_k of $\mathbf{q}$ is the weight of term k in the query. The document and query vectors reside in a t-dimensional vector space, where t is the number of terms in the indexing vocabulary.

Only single terms were used in the present investigation. Document term weights were SMART *lnc* weights, and query term weights were SMART *ltc* weights (Buckley et al., 1995). The formula for the *lnc* document term weight is

$$d_{ik} = \frac{\log(f_{ik}) + 1}{\sqrt{\sum_{j=1}^{t} \left(\log(f_{ij}) + 1\right)^2}} \ , \qquad (1)$$

where f_{ik} is the number of times term k appears in document i. The within-document frequency, f_{ik}, is an indication of the degree to which document i is about the concept represented by term k; thus, as f_{ik} increases, d_{ik} increases. The denominator of Equation 1 is a document-length normalization factor; this normalization keeps short documents from being penalized.

The formula for the *ltc* query term weight is similar to *lnc*:

$$q_k = \frac{\left(\log(f_k) + 1\right) * idf_k}{\sqrt{\sum_{j=1}^{t} \left[\left(\log(f_j) + 1\right) * idf_j\right]^2}} \qquad (2)$$

where f_k is the number of times term k appears in the query, and idf_k is the inverse document frequency (Sparck Jones, 1972) of term k. Like within-document frequency, the within-query frequency, f_k, is an indication of the degree to which the query is about the concept represented by term k; thus, as f_k increases, q_k increases. As in Equation 1, the denominator of Equation 2 is a length normalization factor. Finally, the inverse document frequency, idf_k, is computed by $\log(N/n_k)$ where N is the number of documents in the collection and n_k is the number of documents in the collection that are indexed by term k. The greater the number of documents an index term k represents, the lower the power of k to discriminate documents. Thus, as n_k increases, idf_k decreases and q_k decreases.

SMART created the initial ranking of documents. The documents were ranked in decreasing order of the dot-product $\mathbf{q} \bullet \mathbf{d_i}$, where

$$\mathbf{q} \bullet \mathbf{d_i} = \sum_{k=1}^{t} q_k d_{ik} \,. \tag{3}$$

The dot product of $\mathbf{d_i}$ and $\mathbf{q}$, whose components are defined by Equations 1 and 2, respectively, is the cosine similarity of the two vectors (Salton & McGill, 1983, p. 121).

Relevance Feedback Algorithms

Two relevance feedback models were studied. Run *uncis1* employed the adaptive linear (AL) model (Wong & Yao, 1990). Run *uncis2* employed a probabilistic model, following Robertson and Sparck Jones (1976).

In an effort to produce comparable results, several aspects of the runs were controlled. First, as discussed previously, the top 5000 documents of the SMART initial ranking for a topic served as a condensed collection for relevance feedback operations on that topic. Second, relevance feedback operations ranked documents in decreasing order of $\mathbf{q_m} \bullet \mathbf{d_i}$, where $\mathbf{q_m}$ is the query vector for iteration m. Here, the initial query vector is $\mathbf{q_1}$, the modified query vector after the first set of relevance evaluations is $\mathbf{q_2}$, and so on. Defined by Equation 1, components of the document vector $\mathbf{d_i}$ were held constant for all iterations of both runs. Thus, the advantages of employing within-document frequency and document-length normalization apply to the relevance feedback iterations as well as to the initial ranking. Third, a query was expanded by including all terms in retrieved, relevant documents. (For some of the *uncis1* iterations, all terms in selected nonrelevant documents were included as well.) Some justification for expanding the query by all of the terms in retrieved, relevant documents is provided by Salton and Buckley (1990). They examined a number of different relevance feedback models and several different expansion methods, and found that expanding the query by all terms in relevant documents generally produced better results than other expansion methods studied. However, in contrast to the WSJ collection, the collections studied in Salton and Buckley's experiment consisted of document surrogates (such as abstracts and titles). Haines and Croft (1993), studying the full-text collection, WEST, found that for some expansion methods, adding too many terms to the query caused a substantial decrease in performance. They did not find such a decrease in the document-surrogate collection, CACM. Fourth, query term weights were cumulated as follows:

$$\mathbf{q_{m+1}} = \mathbf{q_m} + \mathbf{x}. \tag{4}$$

The query vector $\mathbf{q_{m+1}}$ for the iteration m+1, is created by the vector sum of the query vector $\mathbf{q_m}$ for the current iteration m and the vector $\mathbf{x}$. The vector $\mathbf{x}$ depends on the algorithm employed by each run. For example, for many of the *uncis1* iterations, $\mathbf{x}$ was the vector sum of all of the *new* relevant documents retrieved by $\mathbf{q_m}$. (Here, a "new" relevant document is one that was not retrieved in a prior iteration.) An advantage of cumulating query term weights is that information from previous iterations contributes to the new query vector. Thus, the query vector does not veer off radically in a new direction due to the influence of new documents retrieved in the latest iteration. Finally, the official run submitted for a topic was the SMART initial ranking if no documents retrieved from this ranking were declared relevant. Otherwise, the ranking resulting from the final set of relevance evaluations was submitted for a topic.

We did not control for the number of iterations or the number of documents assessed in each iteration. All top-ranked documents retrieved in an iteration were assessed for relevance. An iteration (and a search) generally ended when the searcher felt there would be no further benefit to continue to assess documents. Relevance judgments were made independently in *uncis1* and *uncis2* and appear to have undermined all other efforts to produce comparable results; that is, to produce comparisons attributable primarily to the algorithms.

adaptive linear model (run *uncis1*)

Run *uncis1* employed the adaptive linear (AL) model (Bollmann & Wong, 1987; Wong & Yao, 1990; Wong, Yao, & Bollmann, 1988; Wong, Yao, Salton, & Buckley, 1991). The AL model is based on the concept of the preference relation (Fishburn, 1970; Roberts, 1979). The *user preference relation* $\succ$ on a set of documents D is defined as a binary relation on D where for all d, $d' \in D$,

$$d \succ d' \Leftrightarrow \text{the user with a query prefers } d \text{ to } d' \tag{5}$$

(Bollmann & Wong, 1987). An IR model based on the user preference relation allows the use of a multivalued relevance scale—for example, "relevant," "marginally relevant," and "not relevant." However, for TREC-5, only binary relevance distinctions were made.

Wong et al. (1991) assume that $\succ$ on D can be mapped to a corresponding relation on $\mathbf{D}$, the set of document vectors for the documents in D. Thus, for all d, $d' \in D$,

$$d \succ d' \Leftrightarrow \mathbf{d} \succ \mathbf{d'}, \tag{6}$$

where d and d' are represented, respectively, by $\mathbf{d}$ and $\mathbf{d'}$ and where $\succ$ is also used for the relation on $\mathbf{D}$ for the sake of convenience. The relation $\succ$ on $\mathbf{D}$ is defined to be *weakly linear* if there exists a query vector $\mathbf{q}$ such that for all $\mathbf{d}$, $\mathbf{d'} \in \mathbf{D}$,

$$\mathbf{d} \succ \mathbf{d'} \Rightarrow \mathbf{q} \bullet \mathbf{d} > \mathbf{q} \bullet \mathbf{d'} \tag{7}$$

(Wong & Yao, 1990; Wong et al., 1988). Such a query vector is called a *solution vector*. In the context of binary relevance judgments, if $\succ$ on $\mathbf{D}$ is weakly linear, then a solution vector $\mathbf{q}$ will rank all of the relevant documents in the collection before all of the nonrelevant ones.

If $\succ$ on $\mathbf{D}$ is weakly linear, then a solution vector can be found by employing an *error-correction procedure* (Nilsson, 1965, Ch. 4; Wong et al., 1988). However, a user's preferences are only known for the set of documents (the *sample set*) that have been retrieved and assessed for relevance; a document in the sample set may have been retrieved during previous search iterations or the current one. Thus, in the AL model, a solution vector for $\mathbf{S}$ is found, where $\mathbf{S}$ is the set of document vectors corresponding to the sample set *S*. Wong and Yao (1990) assume that, as $\mathbf{S}$ grows larger, the solution vector for $\mathbf{S}$ approaches the solution vector for $\mathbf{D}$. A solution vector for $\mathbf{S}$ can almost always be found; $\succ$ on a set of document vectors is usually weakly linear when the number of terms in the indexing vocabulary is much greater than the number of vectors in the set (Nilsson, pp. 32-35).

For TREC-5, the solution vector for $\mathbf{S}$ was determined by employing the error-correction procedure used by Wong et al. (1991). If $\succ$ on $\mathbf{S}$ is weakly linear, it can be shown that this algorithm will converge to a solution vector (Duda & Hart, 1973, pp. 142-145; Nilsson, 1965, pp. 85-87). The error-correction procedure is an iterative algorithm. During a given cycle *i* of the algorithm, if query vector $\mathbf{q}_{(i)}$ is a solution vector, then the algorithm terminates. If $\mathbf{q}_{(i)}$ is not a solution vector, a new query vector $\mathbf{q}_{(i+1)}$ is created by

$$\mathbf{q}_{(i+1)} = \mathbf{q}_{(i)} + \mathbf{b}_{max} \tag{8}$$

where

$$\mathbf{B} = \{\mathbf{b} = \mathbf{d} - \mathbf{d}' \mid \mathbf{d}, \mathbf{d}' \in \mathbf{S} \text{ and } \mathbf{d} \succ \mathbf{d}'\} \tag{9}$$

and where $\mathbf{b}_{max} \in \mathbf{B}$ such that for all $\mathbf{b} \in \mathbf{B}$,

$$-\mathbf{q}_{(i)} \bullet \mathbf{b}_{max} \geq -\mathbf{q}_{(i)} \bullet \mathbf{b}. \tag{10}$$

As the desired result is $\mathbf{q}_{(i)} \bullet \mathbf{b} > 0$ for all $\mathbf{b} \in \mathbf{B}$, then the quantity $-\mathbf{q}_{(i)} \bullet \mathbf{b}$ can be viewed as a measure of the extent to which $\mathbf{b}$ is in error; $\mathbf{b}_{\max}$ then is that $\mathbf{b}$ that produces the maximum error (Wong et al., 1991). In TREC-5, only binary relevance assessments were made. Accordingly, in the context of binary relevance judgments,

$$\mathbf{q}_{(i+1)} = \mathbf{q}_{(i)} + \mathbf{d}_{worst\_rel} - \mathbf{d}_{best\_nonrel} \tag{11}$$

where the documents in *S* have been ranked by $\mathbf{q}_{(i)}$, $\mathbf{d}_{worst\_rel}$ represents the worst-ranked relevant document, and $\mathbf{d}_{best\_nonrel}$ represents the best-ranked nonrelevant document.

The *starting vector* $\mathbf{q}_{(0)}$ is the initial query vector of the error-correction procedure (Wong et al., 1991). (If the starting vector is a solution vector, then the procedure terminates without creating $\mathbf{q}_{(1)}$.) Wong et al. conducted relevance feedback runs using three different starting vectors, and the best results were for those runs where the starting vector was the user's original query vector plus the vectors for the retrieved, relevant documents. Wong et al. conducted only one feedback iteration; because we conducted more than one feedback iteration, our starting vector consisted of the query vector for the current search iteration plus the vectors for all *new* relevant documents retrieved.

Using binary relevance assessments, Wong et al. (1991) compared the AL model to some other relevance feedback models that had been found to produce good results. The AL model's average precision was slightly superior to that of these other models. The five test collections studied in the Wong et al. experiment consisted of document surrogates (such as abstracts and titles) instead of full-text documents. Therefore, we felt that it would be interesting to evaluate the AL model using the full-text and much larger WSJ collection.

probabilistic model (run *uncis2*)

As in the implementation of the AL model, the query vector at search iteration m+1, $\mathbf{q_{m+1}}$, is the vector sum of the query vector at iteration m, $\mathbf{q_m}$, and the vector $\mathbf{x}$. The complete set of documents evaluated by the searcher in the ranked outcome of iteration m defines the *training set*, and $\mathbf{x}$ includes all terms in relevant documents of this set. The weight of term k in $\mathbf{x}$ is denoted by $(tr)_k$ and is defined by

$$(tr)_k = \log \left[\frac{p_k / (1 - p_k)}{u_k / (1 - u_k)} \right], \qquad (12)$$

where p_k is the probability term k appears in a relevant document of the training set, and u_k is the probability term k appears in a nonrelevant document in the training set (Robertson & Sparck Jones, 1976). Definitions of p_k and u_k can lead to undefined values of $(tr)_k$, when a term appears in all or none of the relevant or nonrelevant documents in the training set, and consequently, computing equations must be used to estimate the probabilities. Because the "conventional" 0.5 formula (Robertson & Sparck Jones) can overestimate term relevance weights, especially when few relevant documents are detected in the training set (Shaw, 1995; van Rijsbergen, Harper, & Porter, 1981; Yu, Buckley, Lam, & Salton, 1983), "alternative" computing equations were used to determine p_k and u_k (Shaw, 1995):

$$p_k = \frac{r_k}{N_r} \left[\begin{array}{c} \dfrac{1}{N_d^{\,2}},\ if\ r_k = 0 \\ 1 - \dfrac{1}{N_d^{\,2}},\ if\ r_k = N_r \end{array} \right] \qquad (13)$$

and

$$u_k = \frac{d_k - r_k}{N_d - N_r} \left[\begin{array}{c} \dfrac{1}{N_d^{\,2}},\ if\ d_k - r_k = 0 \\ 1 - \dfrac{1}{N_d^{\,2}},\ if\ d_k - r_k = N_d - N_r \end{array} \right], \qquad (14)$$

where N_d and N_r are the total number of documents and the total number of relevant documents, respectively, in the training set, and d_k and r_k are the number of documents and the number of relevant documents, respectively, in which term k appears. The alternative computing equations contribute to a high level of retrieval effectiveness in retrospective and predictive tests in a small retrieval test collection (Shaw, 1995, 1996) and are employed here for comparative purposes. In all feedback iterations, documents were ranked according to the cosine similarity of document and query vectors.

Results

The number of feedback iterations for *uncis1* and *uncis2* varied from 0 to 4 and from 0 to 5 respectively. If the searcher declared no document to be relevant in the SMART initial ranking, the number of feedback iterations was 0; for both *uncis1* and *uncis2*, 14 topics out of 50 had no feedback iterations. Considering only those topics with feedback iterations, the average number per topic for *uncis1* and *uncis2* were 2.78 and 2.83 respectively.

For those topics for which the searcher declared no document to be relevant in the SMART initial ranking, the average number of documents evaluated from this ranking for *uncis1* was 68.29. For all other topics, the average number of documents evaluated from the initial ranking was 26.53. The average number of *new* documents evaluated from the first feedback ranking for *uncis1* was 14.47, and this average for subsequent rankings was 8.50. Corresponding data are not recoverable for *uncis2*; however, an attempt was made to examine 30 new documents per iteration. For both *uncis1* and *uncis2*, the actual number of new documents evaluated varied widely.

We have assumed that a high fraction of relevant documents appears among the first 5000 documents ranked by SMART. A comparison of results for TREC-4 and -5 is given in Table 2 and shows that initial rankings in TREC-5 are less reliable than initial rankings in TREC-4, but are quite satisfactory. For 51% of topics with at least one relevant document in TREC-5, 100% of the relevant documents appear among the first 5000 documents, while 67% of topics in TREC-4 meet that standard. Similarly, 73% of TREC-5 topics yield 80% or more of the relevant documents in the initial rankings, while 96% of TREC-4 topics are so productive. Our results in TREC-5 are constrained by a few topics for which high recall is not possible.

Summary statistics for our TREC-5 results are given in Tables 3 and 4. These results are based on the 45 topics with at least one relevant document. Results for a baseline run consisting of the SMART initial ranking are also given.

Table 2: For n = 5000, the number of TREC-4 and TREC-5 ad hoc topics with the value for recall within the stated range.

Recall (*Rec*)	TREC-4 Topics	TREC-5 Topics
$Rec = 1.000$	30	23
$0.900 \leq Rec < 1.000$	7	7
$0.800 \leq Rec < 0.900$	6	3
$0.700 \leq Rec < 0.800$	0	5
$0.600 \leq Rec < 0.700$	0	2
$0.500 \leq Rec < 0.600$	1	2
$0.400 \leq Rec < 0.500$	1	1
$0.300 \leq Rec < 0.400$	0	1
$0.200 \leq Rec < 0.300$	0	0
$0.100 \leq Rec < 0.200$	0	0
$0.000 < Rec < 0.100$	0	0
$Rec = 0.000$	0	1

Table 3: Recall level precision averages and average non-interpolated precision for baseline run, *uncis1*, and *uncis2*.

Recall Level Precision Averages			
Recall	baseline precision	*uncis1* precision	*uncis2* precision
0.00	0.3126	0.5704	0.5156
0.10	0.2589	0.5185	0.4299
0.20	0.2037	0.4403	0.3703
0.30	0.1643	0.3674	0.3270
0.40	0.1282	0.3242	0.2248
0.50	0.1146	0.2904	0.1957
0.60	0.0816	0.1961	0.1389
0.70	0.0696	0.1092	0.0730
0.80	0.0446	0.0706	0.0371
0.90	0.0376	0.0493	0.0209
1.00	0.0318	0.0408	0.0181
Average precision over all relevant documents			
	baseline	*uncis1*	*uncis2*
non-interpolated	0.1172	0.2510	0.1948

Table 4: Document level precision averages and average R-precision for baseline run, *uncis1*, and *uncis2*.

Document Level Averages			
Cutoff	baseline precision	*uncis1* precision	*uncis2* precision
At 5 docs:	0.1556	0.3867	0.3333
At 10 docs:	0.1444	0.3289	0.2844
At 15 docs:	0.1230	0.2785	0.2281
At 20 docs:	0.1111	0.2356	0.1956
At 30 docs:	0.0926	0.1807	0.1459
At 100 docs:	0.0487	0.0824	0.0644
At 200 docs:	0.0333	0.0488	0.0407
At 500 docs:	0.0199	0.0249	0.0215
At 1000 docs:	0.0128	0.0145	0.0135
R-Precision (precision after R (number relevant for a query) documents retrieved)			
	baseline	*uncis1*	*uncis2*
Exact	0.1428	0.2470	0.2123

Discussion

Our objective was to compare the effectiveness of AL and probabilistic models in the context of relevance feedback. With respect to overall (average non-interpolated) precision, performance outcomes for *uncis1* and *uncis2* rank first and fourth, respectively, among official results reported by Category B participants. As both runs were the only Category B runs to incorporate relevance feedback utilizing searcher relevance assessments, their high performance relative to the other runs offers evidence for the effectiveness of this retrieval strategy. However, with overall precision values ranging from 0.20 to 0.25, results are poor for both models in absolute value. A query-level analysis provides some insight into low, official levels of performance.

Table 5 gives performance rankings of topics in decreasing order of average non-interpolated precision for *uncis1*, recall values at n=5000, rank of the first official relevant document in the initial search, overlap of judges' and searchers' relevance assessments, and average non-interpolated precision values for both runs. The table includes 45 topics with at least one official relevant document. Focusing on recall, it can be seen that a high fraction of relevant documents appears in the condensed collection for most topics. Including topic 253 (performance rank 44), for which no officially declared relevant document is found among the first 5000 documents, an average of 87% of relevant documents for a topic appeared in the associated condensed collection. In addition, measurable values of precision are reported for standard recall value 1.00 for both runs, as shown in Table 3. The condensed document collection for each topic does not explain reported performance levels.

Table 5: Factors that may influence performance results. Only the 45 topics with an "official" relevant document are shown.

Performance Rank[1] / Topic #	Recall at $n = 5000$	Rank of first relevant document[2]	*uncis1*		*uncis2*	
			Overlap[3]	Average precision[4]	Overlap[3]	Average precision[4]
1 / 276	1.000	1	0.400	0.7000	0.286	0.6667
2 / 259	1.000	3	0.636	0.6971	0.667	0.7184
3 / 274	0.912	23	1.000	0.6786	1.000	0.5035
4 / 300	0.933	11	0.818	0.6699	0.833	0.5422
5 / 288	1.000	1	0.722	0.6073	0.700	0.4940
6 / 285	1.000	1	0.938	0.6060	0.710	0.5679
7 / 287	1.000	4	0.714	0.5909	0.300	0.1952
8 / 261	0.900	1	0.714	0.5275	0.333	0.3550
9 / 272	1.000	1	0.286	0.4675	0.500	0.4903
10 / 258	1.000	31	0.500	0.4217	0.091	0.0387
11 / 289	0.988	2	0.640	0.4159	0.606	0.3223
12 / 299	1.000	4	0.800	0.4061	0.300	0.3691
13 / 280	1.000	2	0.333	0.3803	0.400	0.3559
14 / 256	1.000	2	0.444	0.3560	0.500	0.5694
15 / 257	0.800	5	1.000	0.3546	0.333	0.1133
16 / 298	1.000	23	0.714	0.3510	0.750	0.2835
17 / 252	1.000	210	0.111	0.3333	0.063	0.1000
18 / 254	0.923	1	0.539	0.3247	0.303	0.1133
19 / 273	0.796	5	0.889	0.3157	0.833	0.2321
20 / 271	0.950	4	0.471	0.2815	0.667	0.2911
21 / 293	1.000	2	1.000	0.2611	0.833	0.3300
22 / 282	0.857	5	0.400	0.2074	0.400	0.2695
23 / 255	0.906	9	0.462	0.2015	0.444	0.1092
24 / 291	0.841	3	0.765	0.2009	0.364	0.0547
25 / 284	1.000	1	0.263	0.1706	0.263	0.1003
26 / 283	1.000	9	0.105	0.1385	0.231	0.1223
27 / 265	1.000	302	0.200	0.1290	0.000	0.0256
28 / 294	0.556	6	0.500	0.1176	0.400	0.0545
29 / 290	0.676	281	0.214	0.1093	0.385	0.0878
30 / 266	1.000	10	0.200	0.1000	0.143	0.1714
31 / 268	0.700	20	0.500	0.0600	0.073	0.0111
32 / 286	0.750	85	0.111	0.0364	0.200	0.0175
33 / 251	0.635	10	None found	0.0233	None found	0.0233
34 / 292	1.000	71	0.000	0.0155	0.000	0.0033
35 / 262	1.000	85	None found	0.0118	None found	0.0118
36 / 269	0.598	34	0.031	0.0066	0.130	0.0309
37 / 264	0.724	26	None found	0.0054	None found	0.0054
38 / 295	1.000	188	None found	0.0053	None found	0.0053
39 / 277	1.000	179	None found	0.0037	None found	0.0037
40 / 260	1.000	387	None found	0.0026	None found	0.0026
41 / 275	0.333	661	None found	0.0005	None found	0.0005
42 / 297	0.731	495	None found	0.0005	None found	0.0005
43 / 270	0.471	733	None found	0.0001	0.000	0.0013
44 / 253	0.000	> 5000	None found	0.0000	None found	0.0000
45 / 267	1.000	1863	0.000	0.0000	None found	0.0000

[1]Performance rank determined by average precision value for *uncis1*.
[2]Rank of the first "official" relevant document in the initial ranking of documents.
[3]Determined by $x/(x + y)$ where x is the number of documents declared relevant by both the official TREC judges and the searcher, and y is the number of documents declared relevant by the searcher but nonrelevant by the judges. The value given for "Overlap" is "None found" if the searcher declared no document to be relevant.
[4]Average non-interpolated precision over all of the relevant documents.

Rank of first relevant document provides insight into the poor performance of some topics. Table 5 shows that performance is generally poor if the rank of the first official relevant document in the SMART initial ranking is high (greater than 60 or so). For most of these topics, the searcher, as to be expected, declared no retrieved document in the initial ranking to be relevant--indicated by "None found" in the overlap columns for *uncis1* and *uncis2*. For such cases, the ineffective initial ranking becomes the official result. Special mention should be made, however, for three topics where the rank of the first relevant document is high: topics 252, 265, and 290, at performance ranks 17, 27, and 29, respectively. For these topics, nonrelevant documents were declared relevant by the searcher and associated relevant documents were eventually retrieved. Such cases suggest a need for a multivalued relevance scale, including the category "marginally relevant" for documents a searcher considers related to a topic but not an "answer" to it.

Results reported in the table also suggest that differences in performance between the two models are often related to differences in relevance assessments of searchers and judges. Consider topic 258, at performance rank 10, for example, where the overlap for *uncis1* and *uncis2* are 0.50 and 0.09, respectively, and average precision values are 0.42 and 0.04, respectively. Clearly, it would be informative to disassociate discrepancies between relevance assessments by searchers and judges from differences in system performance.

To evaluate the influence of "incorrect" relevance assessments, we selected the 31 topics for which the first official relevant document appears at rank 34 or less, used the official list of relevant documents for all evaluations, and employed an experimental design that replicates aspects of our manual procedures. When we considered at least one document to be relevant in our manual assessments of an initial ranking, we carried out two to five feedback iterations, and evaluated about 8 to 30 new documents in each iteration. In experimental analyses presented here, three feedback iterations were employed. Query expansion and calculations of term weights were based on the complete set of retrieved, evaluated documents, which increased by increments of 25 documents after the first 35 were evaluated in the initial ranking. As in official runs, query expansion strategies employed all terms in relevant, retrieved documents for both models and all terms in selected nonrelevant documents for the AL model, and calculations were based on the sample set, defined in the section describing the AL model. Unlike query vectors computed for *uncis2*, query vectors for the probabilistic model were not cumulated from one iteration to the next.

Official results for the 31 topics are indicated in Figure 1 by the bottom two curves, and experimental results are represented by the middle pair of curves. Clearly, experimental results are superior to official results. Average non-interpolated precision increases from 0.34 (*uncis1*) to 0.60 for the AL model and from 0.27 (*uncis2*) to 0.59 for the probabilistic model, when searcher relevance judgments are replaced with official relevance judgments. (Consistently large sample sets may have also contributed to the increase in performance.) The performance advantage of the AL model indicated by official performance levels is negligible when the confounding influence of inaccurate relevance assessments by searchers are removed. The high relative improvement in performance for the probabilistic model in these implementations appears to be related to two factors:

(1) *uncis2* results were diminished by a lower overlap with official relevance assessments than *uncis1*, and

(2) probabilistic results were enhanced by adopting the sample set instead of the training set for term relevance calculations.

Figure 1 also includes retrospective results for both models. Retrospective results are based on the 5000 documents in the condensed collection for each topic and represent high performance limits for both models. As before, both models employ all terms in relevant documents, while the AL model also uses all terms in selected nonrelevant documents to produce a solution vector. High performance limits of the AL model exceed those of the probabilistic model because the solution vector generated by the former model insures that all relevant documents are ranked ahead of any nonrelevant document in the collection. It is informative to compare recall-precision curves based on a realistically limited view of the collection and on relevance assessments of judges (the middle pair of curves) to an unrealistically complete view of the collection (the retrospective view) and official judgments (the top two curves). When our relevance assessments match the judges', the resulting recall-precision curves are closer to "perfection" than to official results, for standard recall values less than or equal to 0.4. The capacity of systems employing relevance feedback to discriminate relevant documents from nonrelevant documents may be greater than official results suggest.

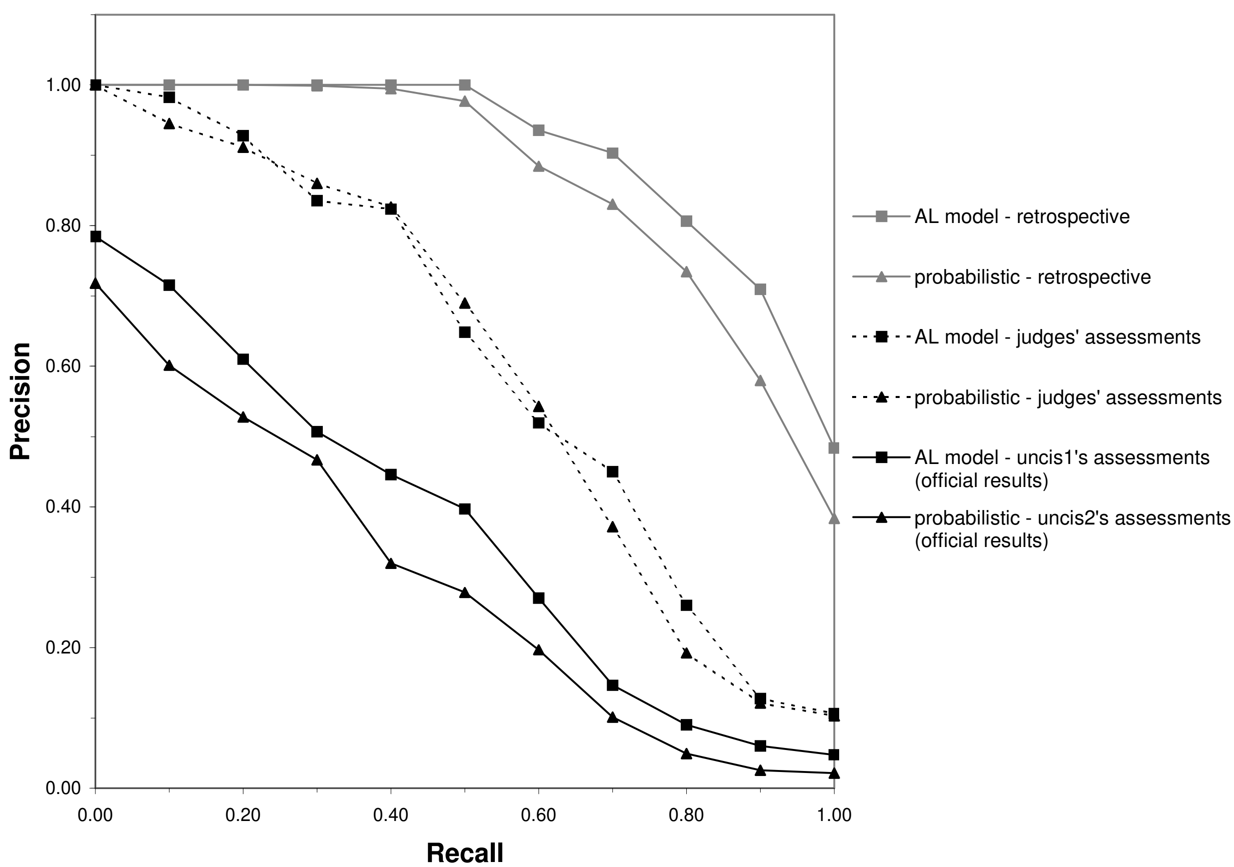

Figure 1 Recall-precision curves. Runs conducted on the 31 topics with at least one relevant document in the top 35 of the SMART initial ranking.

Conclusion

Defined by the first 5000 documents retrieved by an initial search of the WSJ database, the condensed collection for each topic provided a suitable domain for relevance feedback. The condensed collection for each topic included an average of 87% of the relevant documents and contributed to efficient execution of feedback algorithms based on the Adaptive Linear (AL) and probabilistic models. Such a process can lead to one of two outcomes: a satisfactory retrieval result, which is possible in the present context, or a refined query vector, which can be applied to the complete database. We have not yet explored the latter possibility. Implemented as an intelligent client interacting with an end-user and a file server, the two-stage process might offer an efficient and effective model for IR in large distributed databases.

Official results for the AL model are superior to those for the probabilistic model and are first among category B participants. Confounded by distinct, inaccurate relevance assessments by searchers, official results for both models are low in absolute value. Average results for both models increase significantly and become essentially equivalent, when official relevance judgments are used and other influential factors are held constant. Whether the fusion of AL and probabilistic results improve performance remains to be tested. That the AL model allows multivalued relevance assessments may cause it be the model of choice for many applications, if such a capability is favored by end-users, which is likely, and improves performance, which is also likely.

Comparing interactive relevance feedback techniques in TREC, without controlling results for the influence of relevance assessments by searchers, can lead to faulty conclusions, as illustrated here. Declaring a document to be relevant, when judges disagree, can be particularly damaging, if the false positive assessment occurs in an early iteration of the feedback process and aggressive query expansion is employed. Control can be achieved by statistical strategies, as proposed for the TREC-6 interactive track, or by employing official relevance judgments, as illustrated here.

References

Bollmann, P., & Wong, S. K. M. (1987). Adaptive linear information retrieval models. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 157-163.

Boyce, B. (1982). Beyond topicality: A two stage view of relevance and the retrieval process. *Information Processing & Management, 18*, 105-109.

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1995). Automatic query expansion using SMART: TREC 3. In D. K. Harman (Ed.), *Overview of the Third Text REtrieval Conference (TREC-3)* (NIST Spec. Publ. 500-225, pp. 1-19). Washington, DC: U.S. Government Printing Office.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART: TREC 4. In D. K. Harman (Ed.), *The Fourth Text REtrieval Conference (TREC-4)* (NIST Spec. Publ. 500-236, pp. 25-48). Washington, DC: U.S. Government Printing Office.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: John Wiley & Sons.

Fishburn, P. C. (1970). *Utility theory for decision making*. New York: John Wiley & Sons.

Haines, D., & Croft, W. B. (1993). Relevance feedback and inference networks. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2-11.

Nilsson, N. J. (1965). *Learning machines: Foundations of trainable pattern-classifying systems*. New York: McGraw-Hill.

Roberts, F. S. (1979). *Measurement theory with applications to decisionmaking, utility, and the social sciences*. Reading, MA: Addison-Wesley.

Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, *27*, 129-146.

Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M., & Payne, A. (1996). Okapi at TREC-4. In D. K. Harman (Ed.), *The Fourth Text REtrieval Conference (TREC-4)* (NIST Spec. Publ. 500-236, pp. 73-96). Washington, DC: U.S. Government Printing Office.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, *41*, 288-297.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Shalini, R. (1993). "Citation profiles" to improve relevance in a two-stage retrieval system: A proposal. *Information Processing & Management, 29, 463-470.*

Shaw, W.M., Jr. (1995). Term-relevance computations and perfect retrieval performance. *Information Processing & Management, 31,* 491-498.

Shaw, W.M., Jr. (1996). [Letter to the editor]. *Information Processing & Management, 32*, 636-637.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*, 11-21.

van Rijsbergen, C.J., Harper, D.J., & Porter, M.F. (1981). The selection of good search terms. *Information Processing & Management, 17,* 77-91.

Wong, S. K. M., & Yao, Y. Y. (1990). Query formulation in linear retrieval models. *Journal of the American Society for Information Science*, *41*, 334-341.

Wong, S. K. M., Yao, Y. Y., & Bollmann, P. (1988). Linear structure in information retrieval. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 219-232.

Wong, S. K. M., Yao, Y. Y., Salton, G., & Buckley, C. (1991). Evaluation of an adaptive linear model. *Journal of the American Society for Information Science*, *42*, 723-730.

Yu, C.T., Buckley, C., Lam, K., & Salton, G. (1983). A generalized term dependence model in information retrieval. *Information Technology: Research and Development, 2,* 129-154.