Relevance Feedback With Too Much Data

James Allan
allan@cs.umass.edu

Department of Computer Science
University of Massachusetts
Amherst, MA 01003-4610

Abstract

Modern text collections often contain large documents that span several subject areas. Such documents are problematic for relevance feedback since inappropriate terms can easily be chosen. This study explores the highly effective approach of feeding back passages of large documents. A less-expensive method that discards long documents is also reviewed and found to be effective if there are enough relevant documents. A hybrid approach that feeds back short documents and passages of long documents may be the best compromise.

1 Introduction

As the amount of on-line text has increased, so has the size of individual documents in those collections. Information retrieval methods that could easily be applied to the full text of abstracts or short documents are sometimes less effective or prohibitively expensive for large documents. This problem has led to a resurgence of interest in techniques for handling large texts, including passage retrieval, theme identification, document summarization, and so on.

Most work in this area has been done in the "ad-hoc" setting, where retrieval is performed in the absence of known relevant documents. Surprisingly, little work has been done toward applying the same techniques to the information filtering or routing environment, where a collection of documents has been judged for relevance.

After an outline in Section 2 of the methods and environment used in this study, Section 3 examines the value of using passages of long documents for feedback, even in the absence of information about which passage of a relevant document contains the relevant information. The issue of passage length is also examined. Section 4 explores the intriguing idea of totally ignoring large documents, possibly saving a great deal of computational expense compared to passage handling. Section 5 proposes and evaluates a hybrid approach which uses passages of extremely large documents, and the entire text of short and medium-sized documents.

All work in this study was performed using Inquery, a probabilistic information retrieval system based on an inference net model.[Tur90]

Appears in Proceedings of SIGIR '95, pp. 337-343.
July 1995, Seattle, Washington, USA.

	Relevance judgements			
		Per query		
	Total	Avg	Min	Max
Tipster1&2 (training)	16386	327.7	40	894
Tipster3 (eval)	10981	219.6	4	751

Table 2: Query set information

2 Experimental method

In relevance feedback, a query is combined with a set of documents whose relevance to the query is known, and a new—presumably more useful—query is created. If the new query is applied to the same collection of documents from which the training data was drawn, evaluation becomes somewhat complicated.[Sal89] To eliminate most of the difficulties associated with the evaluation of relevance feedback methods, this study uses one collection for the training (feedback) and a *different* collection for evaluating the modified queries. The training collection was the 2Gb collection of documents that comprise the Tipster 1 and 2 datasets. The evaluation was performed on the 1Gb of documents in the Tipster 3 dataset.[Har92] TREC topics 51 through 100 were used for the study.[Har93] Table 1 gives detailed information about the collections, and Table 2 presents some statistics regarding the queries' relationship to the collections.

The queries used in training were created from the TREC topics using an automatic process that modifies the original topic terms by identifying phrases and replacing domain-dependent features with meta-terms.[CCB93] These queries were applied to the training collection and the top n documents were retrieved, where n varied depending on the experiment.

In some experiments, the highest-ranked passage of a document was used for feedback in place of the entire document. Intuition suggests that since writers group related thoughts into a paragraph, passages should be based upon paragraphs. Attempts have been made to use passages derived from text sections,[Wil94] from clusters of paragraph,[RWJ+95] and from arbitrarily long strings of related sentences.[HP93] However, experiments have demonstrated that fixed-size passages are at least as effective—and marginally more efficient—than their varying counterparts.[Cal94, BSAS95]

Passage sizes were thus fixed at some length p for each experiment. The first passage began at the first term of the document that matched a query term, and ended p terms after that. Subsequent passages began at intervals of $p/2$ from that starting point. For ex-

Collection	Numdocs	Size	Contents
Tipster1&2	741,562	2.2Gb	Associated Press 1988-9, Department of Energy abstracts, Federal Register 1988-9, Wall Street Journal 1987-91, Ziff-Davis *Computer-Select* articles
Tipster3	336,310	1.2Gb	AP 1990, Patent claims 1993, San Jose Mercury News 1991 Ziff-Davis

Table 1: Collection statistics

ample, if $p = 200$ and the first term matching the query is the 33rd in the document, passages would start at terms 33, 133, 233, and so on. The same technique used to rank documents can then be applied to the passages of a document, and the best-ranked passage can be selected.[Cal94]

The r relevant documents (or their passages) within the top n retrieved were used for feedback. The method employed has been discussed in detail elsewhere,[HC93] but is summarized here. All terms in the selected relevant documents *that are not in the query* are identified and ranked by:

$$w_k = r_k \cdot \frac{\log((N + 0.5)/n_k)}{\log(N + 1)}$$

where r_k is the number of relevant documents containing term k, N is the total number of documents in the collection, and n_k is the total number of documents that contain the term.[1] (Note that this weight is very similar to a "tf-idf" weighting scheme.) The top $t = \min(3 + 2r, 300)$ terms from that ranked list are chosen and added to the query. Original query terms are given assigned a scaling factor of 1.0; added terms, of 0.3. Finally, the original and new query terms are re-weighted based upon their frequency in the collection (*tf*) according to:

$$new\_wt_k = scaling\_factor_k \cdot tf_k \cdot \frac{\log((N + 0.5)/n_k)}{\log(N + 1)}$$

Phrases and other complex operators are re-weighted similarly.

Finally, the resulting modified queries were applied to the evaluation collection for measuring their effectiveness. Because the evaluation data is distinct from the training data, no "rank freezing" or "residual collection" issues arise. 11-point average precision is used as the basis of evaluation throughout this study.

3 Passage feedback

In traditional relevance feedback environments, a query is modified based upon judgements of the relevance of some number of retrieved documents. Typically the terms occurring in relevant (and sometimes non-relevant) documents are extracted, weighted, and added to the query: terms that occur frequently in relevant documents should help improve retrieval performance. Some collections, however, contain very large documents with wide-ranging discourse. In such documents, selecting terms without regard to position in the document may be a mistake: the relevant portion of the document might be quite small, and the terms should ideally be chosen from only that region.

It is difficult to evaluate the effectiveness of passage selection. Standard retrieval test collections, unfortunately, do not include relevance assessments at the passage level: only entire documents are judged.[2] We believe, nonetheless, that it is important to consider feedback methods that utilize passages of documents rather than entire documents. We can evaluate the effectiveness of doing so by comparing the resulting queries against those generated using full document feedback.

Passages have received quite a bit of attention in recent research, but that work falls primarily into two categories:

1. Many attempts have been made to adjust the ranking of documents with respect to a query based upon the best passage (or passages) of the document: the belief or similarity of a document is replaced or modified by the belief or similarity of one of its passages. [RWJ+95, Cal94, KMS95, SAB93, Wil94] In some cases, the passage similarity has been used as a precision-enhancing filter.[SB91, SAB93] These efforts have typically been successful in improving the effectiveness of retrieval.

2. Alternately, top ranking documents are *assumed* to be "relevant" and they or their passages are used in feedback to generate a new query that is automatically re-applied to the collection. This approach was used, for example, to find terms that co-occurred in the set of retrieved documents so they could be used for query expansion.[AF77] More recently the approach has been applied to more general feedback methods.[KG95, TT95] Note that this technique is very much like "standard" relevance feedback, except that the relevance of documents is assumed, not known.

Oddly, passages have rarely been used for query expansion in a true relevance feedback or routing setting. One exception is [TT95] that discusses passage feedback in terms similar to part of this study, but not as broadly or in as much detail.

3.1 Optimum passage size

We first set out to verify past experience regarding an optimum passage length. Earlier research had suggested that when fixed-length passages are used, anywhere from 200 to 300 words is a good choice for a variety of collections.[Cal94] Other researchers have used passages exceeding 500 words.[KG95]

We are interested in the effect of passage retrieval for interactive uses of a system (small values of n) as well as the routing environment (large n). We therefore ran a set of experiments that varied the number of documents retrieved (hence the number of relevant documents available for feedback) and the choice of passage size from which words were selected for feedback. Table 3 shows the average 11-point precision of 48 experiments retrieving from 5 to 1000 documents and using passage sizes of 50 to 1000 words.

[1] If $r_k = n_k$, then the weight is forced to zero since no other documents contain the term, so it cannot be useful in retrieving additional relevant documents.

[2] One of the few collections that attempted to address that point is the TREC-2 collection which recorded the point in each document at which the relevance assessors decided that the document was (or was not) relevant. Unfortunately, it is not possible to determine if that location in the text was the only relevant one, whether it contained the last bit of information needed to judge relevance, or whether the assessor merely became tired at that point and made a decision. Some work has been done on the same collection to create more useful passage-level relevance judgements.[Wil94]

Passage size	n (number of documents retrieved) 5	25	50	100	1000
full	30.0	30.6	30.2	29.7	29.6
50	30.0	31.6	32.4	32.6	32.3
	-0.0	+3.5	+7.2	+9.8	+9.0
100	29.9	32.0	32.5	32.4	32.2
	-0.3	+4.8	+7.4	+9.0	+8.8
150	29.9	31.9	32.5	32.4	32.2
	-0.4	+4.6	+7.7	+9.2	+8.7
200	30.0	31.8	32.5	32.5	32.0
	+0.1	+4.1	+7.5	+9.4	+7.9
300	30.4	32.2	32.7	32.4	32.1
	+1.3	+5.5	+8.2	+9.3	+8.2
500	30.2	32.1	32.5	32.3	32.1
	+0.6	+5.0	+7.4	+8.7	+8.4
800	p30.3	31.9	32.4	32.2	32.2
	+1.0	+4.3	+7.2	+8.5	+8.6
1000	30.0	31.8	32.3	32.1	32.0
	+0.1	+4.2	+7.0	+8.1	+8.0

Table 3: Feedback passage size vs. number retrieved (11-pt average recall/precision; "full" is baseline for comparisons)

The top row of the table gives the precision for full-document feedback at each of the values of n. The $+/-$ figure under each precision measure indicates the improvement (or drop) in average precision as compared to full-document feedback at that value of n. The row labelled "50" for example is the result of applying feedback using terms from the best 50-word passage of each relevant document in the top n retrieved. At 100 documents retrieved, feeding back the best 50-word passages yielded an average precision of 32.6%, a 9.8% improvement over the full-document precision of 29.7%.[3]

The results support earlier work indicating that a passage size of 200-300 words works well.

3.2 Varying number of documents retrieved

Table 3 shows an interesting change in effectiveness as the number of documents varies, but it is difficult to see. The effect is explored in Table 4 where additional values of n are added to highlight the changes, but is most clearly visible in the graph of Figure 1. The bottom line on the graph depicts the effectiveness of full-document feedback; the picture highlights the increase in effectiveness due to using passages.

Figure 1(b) is an enlarged view of the left half of the graph. In this figure it is clear that passage retrieval is only marginally useful when few documents are retrieved, but that it very quickly becomes valuable. The effectiveness peaks around $n = 40$ and slowly drops as n increases.

Figure 1(a) shows an unusual effect at when all documents in the collection are retrieved. Although effectiveness of passage retrieval was dropping as the number of documents grew, it shows a sharp climb when *all* relevant documents are included. We believe this is an artifact of the way relevance assessments for this collection were acquired. Most (61%) of the 15,868 relevant documents (over all 50 queries) occurred within the top 1000 retrieved documents, but over 200 of them were not found until after rank 100,000. (One query found 12 of its relevant documents after rank 300,000.) Such documents were probably retrieved for judging by quite different retrieval systems. Once the query has been modified by including those documents, Inquery is implicitly taking advantage of those other systems' capabilities. The effect, therefore, is similar to the improvement resulting from "data fusion" of differing retrieval methods.[FS93] We conjecture that the effect does not occur in full-document feedback because the extraneous text in the longer documents hurts more than the "data fusion" helps.

4 Discarding big documents

It has seemingly become standard to handle large, presumably discursive documents by choosing an appropriate way to divide them into passages. But an alternative approach is to simply discard such documents altogether, potentially creating a substantial savings in retrieval time (passage handling is quite expensive in some systems). Such a method also has the advantage of making moot the question of whether passage retrieval is in fact selecting the relevant passage.

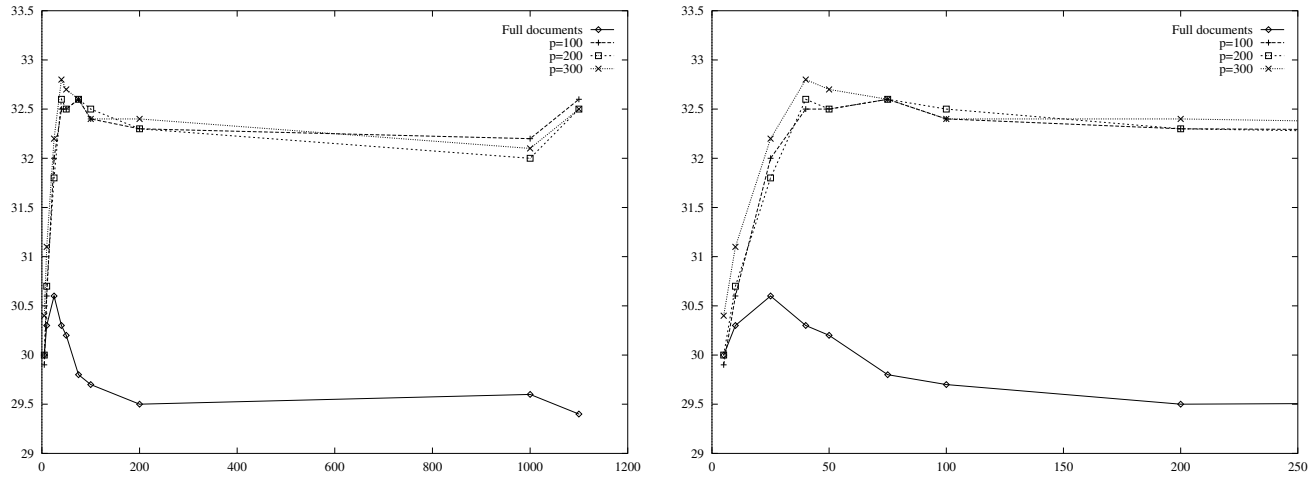
This approach was taken in [KG95] where documents with more than 160 unique non-stopword terms were considered "too big" for feedback. These results of discarding such longer documents were encouraging, but they refer only to the massive feedback (large n) routing environment of TREC. What is the impact of discarding documents in smaller sets of relevant documents?

For this study, we measured the size of documents by the number of non-stopword terms that occur in the document: a term that

[3] Full-document feedback is always an improvement over no query modification at all—*i.e.*, over applying the original queries to the evaluation data. Full-document feedback typically results in a 8-10% improvement over the original query.

Passage size	n (number of documents retrieved) 5	10	25	40	50	75	100	200	1000	all
full	30.0	30.3	30.6	30.3	30.2	29.8	29.7	29.5	29.6	29.4
100	29.9	30.6	32.0	32.5	32.5	32.6	32.4	32.3	32.2	32.6
	-0.3	+0.8	+4.8	+7.4	+7.4	+9.4	+9.0	+9.2	+8.8	+10.8
200	30.0	30.7	31.8	32.6	32.5	32.6	32.5	32.3	32.0	32.5
	+0.1	+1.1	+4.1	+7.4	+7.5	+9.2	+9.4	+9.3	+7.9	+10.6
300	30.4	31.1	32.2	32.8	32.7	32.6	32.4	32.4	32.1	32.5
	+1.3	+2.6	+5.5	+8.1	+8.2	+9.1	+9.3	+9.6	+8.2	+10.7

Table 4: Passage size vs. more values of n (11-pt average recall/precision; "full" is baseline for comparisons)



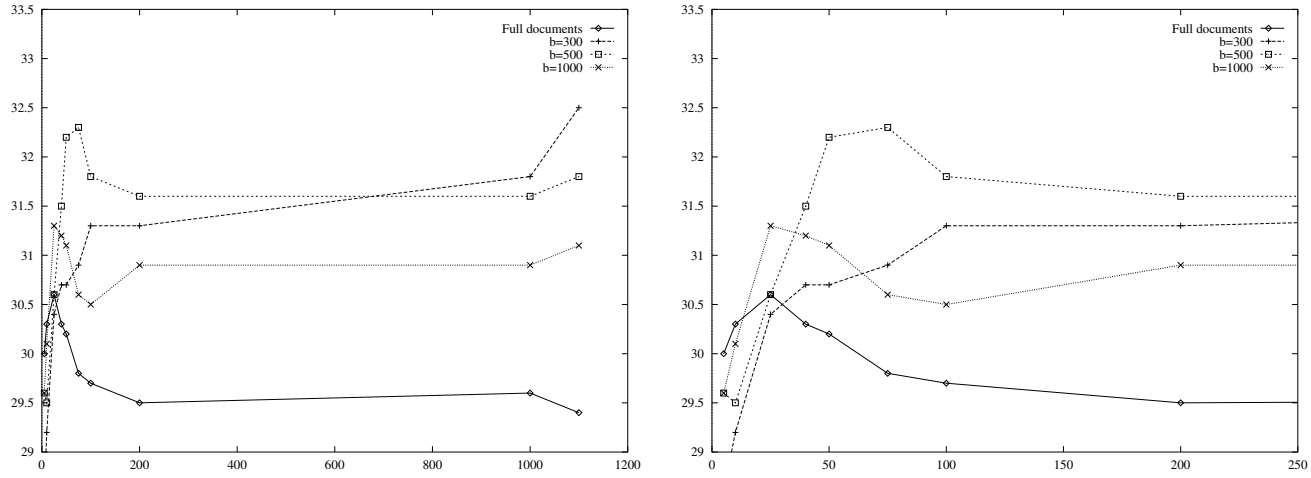
(a) All values of n

(b) Close-up of left side of graph

Figure 1: Passage size vs. number retrieved (from Table 4). Vertical axis is the 11-pt average precision/recall. Each line represents a different passage size.

b	n (number of documents retrieved) 5	10	25	40	50	75	100	200	1000	all
full	30.0	30.3	30.6	30.3	30.2	29.8	29.7	29.5	29.6	29.4
300	28.6	29.2	30.4	30.7	30.7	30.9	31.3	31.3	31.8	32.5
	-4.6	-3.9	-0.5	+1.4	+1.5	+3.7	+5.6	+5.9	+7.3	+10.7
500	29.6	29.5	30.6	31.5	32.2	32.3	31.8	31.6	31.6	31.8
	-1.6	-2.8	+0.3	+3.8	+6.6	+8.1	+7.2	+6.8	+6.7	+8.1
1000	29.6	30.1	31.3	31.2	31.1	30.6	30.5	30.9	30.9	31.1
	-1.4	-0.7	+2.5	+2.8	+2.8	+2.6	+2.6	+4.6	+4.4	+5.7

Table 5: Discarding documents bigger than b words (11-pt average recall/precision; "full" is baseline for comparisons)



(a) All values of n

(b) Close-up of left side of graph

Figure 2: Discarding big documents (from Table 5). Vertical axis is the 11-pt average precision/recall. Each line represents a different passage size.

	n (number retrieved)				
b	5	50	100	1000	total
all	3.5	27.8	49.2	193.8	317.4
300	2.7	15.4	31.8	113.6	176.1
500	1.5	8.9	14.8	49.3	78.8
1000	0.5	2.4	3.7	12.5	23.3

Table 6: Number of relevant documents with more than b words, in top n retrieved

occurs multiple times is counted each time. Since one goal of discarding large documents is to minimize processing, the "largeness" criterion should require minimal processing. (With that in mind, the storage space of the unprocessed document might be an ideal measure, but most modern documents contain a great deal of markup that could cause misleading size estimates.) Table 5 shows the effect that discarding documents has on average precision. As before, the $+/-$ figures reflect the change relative to full-document feedback, the top row of the table.

The negative numbers at the left of the table make it clear that discarding large documents is not necessarily a good approach for small values of n. Figure 2(b) makes that point even more blatantly. For small values of n, discarding documents depresses average precision substantially. However, after roughly 25 documents are retrieved for consideration, full-document feedback is out-performed by discarding large documents (though the "discard" approaches never perform as well as the passage feedback approaches of Figure 1). As n increases there are more relevant documents retrieved and eventually the negative effect of discarding relevant documents is more than out-weighed by the positive effect of dropping the non-relevant material in long documents that are otherwise relevant.

It is not surprising that discarding large documents is counter-productive at small values of n. Consider Table 6 where the average number of relevant documents of varying sizes is shown for different values of n. There are on average 3.5 relevant documents per query in the top 5 retrieved documents, but 2.7 of those are over 300 words, meaning that only 0.8 relevant documents can be used for feedback if documents longer than 300 words are discarded: some queries necessarily get no feedback at all! The document size restriction effectively disables feedback when only a few documents are retrieved. In this training collection, over half of the relevant documents are larger than 300 words at all values of n. Setting the threshold at 500 words substantially increases the number of documents available for feedback. A 1000-word cap is useful, but too many large documents are still available and their inclusion depresses the effectiveness: the cap excludes too few documents to be useful.

Discarding large documents is useful at larger values of n, but it is less clear what is an appropriate definition of "big". Figure 2(b) suggests that documents can be safely ignored if they are larger than 500 terms. However, Figure 2(a) shows that as the number of relevant documents increases, the value of small (300 word) documents increases steadily until it overtakes the larger, 500-word documents. (The "data fusion" effect is noticed here just as it was with passage feedback.)

5 Passages of big documents

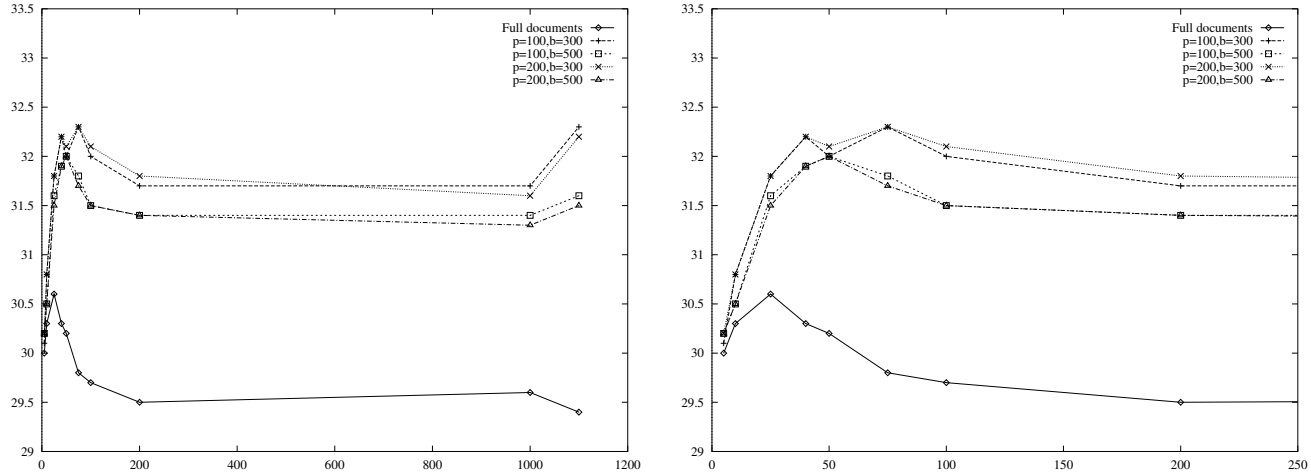
Passage feedback provides a greater improvement in average precision than does discarding long documents, particularly at low values of n. However, when efficiency is as important a consideration as effectiveness, the extra processing required of passage retrieval may be untenable. It is thus reasonable to consider a compromise system: perform passage feedback on "large" documents as necessary, and use the entire text of other relevant documents.

Table 7 shows the result of such a combined run. Full document feedback is used for all documents with fewer than b words, but instead of discarding a remaining "large" document, the best passage of length p is found and its terms are used. (The results of Section 4 suggest that this passage breakdown need be done only if fewer than 10 or so smaller relevant documents are otherwise available.)

Figure 3 is a graph of the combined run information. It has a shape similar to the passage feedback of Figure 1, but the precision values are lower. In fact, the effectiveness of this method is very

		n (number of documents retrieved)									
p	b	5	10	25	40	50	75	100	200	1000	all
full		30.0	30.3	30.6	30.3	30.2	29.8	29.7	29.5	29.6	29.4
100	300	30.1	30.8	31.8	32.2	32.0	32.3	32.0	31.7	31.7	32.3
		+0.2	+1.5	+4.2	+6.3	+5.9	+8.2	+7.8	+7.4	+7.1	+10.0
100	500	30.2	30.5	31.6	31.9	32.0	31.8	31.5	31.4	31.4	31.6
		+0.7	+0.5	+3.3	+5.4	+6.1	+6.5	+6.1	+6.4	+6.0	+7.4
200	300	30.2	30.8	31.8	32.2	32.1	32.3	32.1	31.8	31.6	32.2
		+0.4	+1.6	+4.0	+6.4	+6.4	+8.1	+8.2	+7.6	+6.5	+9.7
200	500	30.2	30.5	31.5	31.9	32.0	31.7	31.5	31.4	31.3	31.5
		+0.7	+0.4	+3.1	+5.4	+5.8	+6.3	+6.1	+6.3	+5.8	+7.1

Table 7: Feeding back passages of length p if document is more than b words long (11-pt average recall/precision; "full" is baseline for comparisons



(a) All values of n

(b) Close-up of left side of graph

Figure 3: Graph of combined runs (from Table 7). Vertical axis is the 11-pt average precision/recall. Each line represents a different passage size.

similar to that of discarding large documents (Figure 2), but the failure of that method at small values of n has been corrected.

6 Conclusion

This study clearly supports the hypothesis that large documents contain information that is debilitating to feedback. Trimming large documents by selecting a good passage has a marked impact on effectiveness. The less computationally expensive approach of discarding large documents altogether is also quite valuable, provided enough "small" relevant documents exist to make feedback possible. A hybrid approach where passages are used only when necessary may prove the most useful—particularly for systems where passage-level operations are expensive or otherwise inappropriate.

We are interested in continuing this work by investigating alternate feedback algorithms. Massive query expansion and negative feedback (of non-relevant documents) have been found useful in the routing setting.[BSAS95] We believe that ultimately a feedback algorithm must adjust its approach based on the number of relevant and non-relevant documents of particular sizes. Our goal is a single algorithm that automatically adapts to the environment rather than a suite of algorithms that a user must decide upon.

Acknowledgements

I am indebted to James Callan and Bruce Croft for their suggestion of this investigation, for their advice during its progress, and for their aid in my learning the Inquery system.

This research was supported in part by the Center for Intelligent Information Retrieval at the University of Massachusetts. This material is also based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsor.

References

[AF77] R. Attar and A. S. Fraenkel. Local feedback in full-text retrieval systems. *Journal of the Association for Computing Machinery*, 24(3):397–417, July 1977.

[BSAS95] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic query expansion using SMART : TREC 3. In *Third Text REtrieval Conference (TREC-3)*, 1995. In press.

[Cal94] James P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310, Dublin, 1994.

[CCB93] Bruce Croft, James Callan, and John Broglio. TREC-2 routing and ad-hoc retrieval evaluation using the INQUERY system. In *Second Text REtrieval Conference (TREC-2)*, pages 75–83, 1993. National Institute of Standards and Technology Special Publication 500-215.

[FS93] Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In *Second Text REtrieval Conference (TREC-2)*, pages 243–252, 1993. National Institute of Standards and Technology Special Publication 500-215.

[Har92] Donna Harman. The DARPA TIPSTER project. *SIGIR Forum*, 26(2):26–28, Fall 1992.

[Har93] Donna Harman. Overview of the second Text REtrieval Conference TREC-2. In *Second Text REtrieval Conference (TREC-2)*, pages 1–20, 1993. National Institute of Standards and Technology Special Publication 500-215.

[HC93] David Haines and W. Bruce Croft. Relevance feedback and inference networks. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–11, Pittsburgh, 1993.

[HP93] Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 59–68, Pittsburgh, 1993.

[KG95] K. L. Kwok and L. Grunfeld. TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In *Third Text REtrieval Conference (TREC-3)*, 1995. In press.

[KMS95] Daniel Knaus, Elke Mittendorf, and Peter Schäuble. Improving a basic retrieval method by links and passage level evidence. In *Third Text REtrieval Conference (TREC-3)*, 1995. In press.

[RWJ+95] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Third Text REtrieval Conference (TREC-3)*, 1995. In press.

[SAB93] Gerard Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–58, Pittsburgh, 1993.

[Sal89] Gerard Salton. *Automatic text processing*. Addison-Wesley, 1989.

[SB91] Gerard Salton and Chris Buckley. Automatic text structuring and retrieval – experiments in automatic encyclopedia searching. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, Pittsburgh, 1991.

[TT95] Paul Thompson and Howard Turtle. TREC-3 ad hoc retrieval and routing experiments using the WIN system. In *Third Text REtrieval Conference (TREC-3)*, 1995. In press.

[Tur90] Howard R. Turtle. *Inference networks for document retrieval*. PhD thesis, University of Massachusetts, Amherst, October 1990. Also technical report 90-92.

[Wil94] Ross Wilkinson. Effective retrieval of structured documents. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 311–317, Dublin, 1994.