To appear in *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July, 1994, Dublin, Ireland.

Passage-Level Evidence in Document Retrieval

James P. Callan
callan@cs.umass.edu

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

Abstract

The increasing lengths of documents in full-text collections encourages renewed interest in the ranking and retrieval of document passages. Past research showed that evidence from passages can improve retrieval results, but it also raised questions about how passages are defined, how they can be ranked efficiently, and what is their proper role in long, structured documents.

This paper reports on experiments with passages in INQUERY, a probabilistic information retrieval system. Experiments were conducted with passages based on paragraphs, and with passages based on text windows of various sizes. Experimental results are given for three homogeneous and two heterogeneous document collections, ranging in size from three megabytes to two gigabytes.

1 Introduction

It is sometimes better to apply retrieval algorithms to portions of a document text than to all of the text. If each portion of text, or *passage*, is ranked, an interface can quickly direct a user to the relevant information in a document [11; 12]. Whether this feature is necessary or merely helpful was unclear from research using short and medium length documents. Recent work using long documents with complex internal structure suggests that passage-level evidence is an important part of document retrieval [14; 13; 15; 10; 6].

Long documents, documents with complex structure, and even short documents summarizing many subjects, are a challenge for algorithms that do not distinguish where in a document the text matches a query. If the algorithm cannot distinguish a few matches scattered across a document from a dense region of matches, it may have difficulty retrieving long documents and newswire "news summaries".

One can view information retrieval as a task of retrieving passages instead of documents [11]. During indexing, a document is divided into passages, perhaps with links maintained between its components, and each passage is stored as if it were a distinct entity [10; 9]. This approach has the advantage of simplicity, because no new algorithms are necessary. However, it creates a problem of how to recognize a small portion of relevant text that spans two passages [15]. It may also be unsuitable for very long queries, because chances are lower that short passages can match many of the query terms.

Information retrieval can also be viewed as a task of retrieving documents that have a hierarchical internal structure [14; 10]. For example, documents may contain sections, paragraphs and sentences. Each element of the document structure is a source of evidence that can be used in retrieval. Document-level matches are *global* evidence, while sentence-level matches are very *local.* Combining evidence from different levels of a document's structure may provide more accurate retrieval than will evidence from any single level [14].

The types of passages explored by researchers can be grouped into three classes: discourse, semantic, and window. *Discourse* passages are based upon textual discourse units (e.g. sentences, paragraphs and sections). *Semantic* passages are based upon the subject or content of the text (e.g. *TextTiling* [9]). *Window* passages are based upon a number of words.

One might expect discourse passages to be the most effective, because discourse boundaries organize material by content. However, discourse passages also require more consistency from writers than do semantic or window passages. If writers are sloppy or rushed, or if discourse boundaries are introduced for visual presentation, then dividing documents by textual discourse boundaries may be inappropriate. Discourse passages have been found to work well with highly structured and edited encyclopedia text [14; 13]. However, in experiments with the TIPSTER collection [8], which contains short, long, structured, and news summary documents, discourse passages sometimes worked well and sometimes did not [2; 10].

In spite of the intuitive appeal of passages, and in spite of the evidence in their favor, it remains unclear how important they are and how to use them effectively. What *is* clear is that passages present a system designer with a variety of challenging issues, ranging from how passages are defined, stored and retrieved to how they are combined with other forms of evidence. There are many more passages than documents, so in general they slow retrieval. How to use this source of evidence, and how to use it efficiently, are nontrivial problems.

This paper describes progress in adding passages to INQUERY, a probabilistic information retrieval system based upon a Bayesian inference network model [16; 4]. Two types of passages, discourse and window, are investigated. The issue of how best to use passage-level evidence, alone or in combination with document-level evidence, is also addressed. Finally, the issue of efficiently ranking passages is discussed briefly.

The work described in this paper differs from earlier work with INQUERY in which "paragraph-level" constraints were expressed with a proximity operator [3; 5]. The advantages of that approach were its precision, efficiency, and simplicity. The disadvantage was the all-or-nothing nature of the operator, which restricted its use to simple constraints formed manually. In the work reported here, a complex structured query can be applied to passages in exactly the same way it can be applied to documents. The principal difference is the amount of text involved.

2 Ranking Passages in INQUERY

Inference networks are an effective formalism for reasoning about multiple sources of uncertain evidence, which makes them particularly suited for IR. Previous research showed that INQUERY's performance, as measured by precision and recall, generally improved with the amount of evidence available. This result held for collections of short documents [16], a collection of large documents [7], and a heterogeneous collection [1].

INQUERY's effectiveness at using multiple sources of evidence is consistent with the view of IR as a task of retrieving structured documents. Multiple sources of evidence can be obtained by applying a query to various levels of a document, or by applying different queries to different levels. This approach requires adding to the query language operators that control the scope of a query or query fragment. In the work reported here, the INQUERY query language was extended slightly, to include a new operator that restricted the scope of its argument (a query or query fragment) to passages.

The implementation of a #passage scope operator required two restrictions on query formulation. The restrictions were 1) a slightly simplified phrase model, and 2) use of the synonym operator in place of the probabilistic OR operator. These restrictions simplified implementation, but are not inherent in the approach taken.

Two experiments investigated whether these restrictions cause any improvements in document retrieval that might be attributed mistakenly to use of passage-level evidence. Both experiments used complete and restricted query sets to retrieve documents from one volume of the TIPSTER collection (about a gigabyte text). TREC judgements were used to assess relevance. In both experiments, the restrictions caused a small loss in the average precision of document retrieval (−2.7% and −3.6%). The results suggest that relaxing these restrictions on query formulation would slightly improve the passage-level results presented below.

If passages are viewed as evidence for *document* retrieval, it makes little sense to store the score or rank of each passage in each document. The amount of memory required to do so for a collection like TIPSTER would be extremely large. Instead, it is sufficient to record the contribution of its passages to a document's score, and then to discard them. No extra memory is required. Preliminary experiments suggested that the best results could be obtained by recording only the contribution provided by a document's best passage. Others have found that adding the scores of several passages is effective [9; 17], but that was not the case in our experiments.

Ignorance of how best to divide documents into passages discouraged the use of special indices. Instead, experiments were conducted by reading the usual document-level indices and dividing documents into passages "on-the-fly" during retrieval. This approach requires only a very small index of discourse boundaries, but does slightly increase the cost of retrieval.

Three different approaches were taken to determining passage size and location: paragraphs, bounded paragraphs, and fixed-length windows. Paragraphs were recognized by heuristic rules (e.g. by indentation). Bounded paragraphs were paragraphs that were constrained to be greater than a minimum number of words and less than a maximum number of words. The sizes of fixed-length windows were specified as query-language parameters. Each approach is described in more detail below.

2.1 Paragraph Passages

A series of experiments was conducted with INQUERY to test discourse passages in the TIPSTER collection. The boundaries between paragraphs are not identified explicitly in the TIPSTER collection, so they were recognized automatically during indexing by a set of collection-specific heuristics. In preliminary experiments, paragraph boundaries performed poorly on collections of short and medium length

Table 1. TIPSTER Federal Register volume 1, selected queries, TREC-1 judgements. Paragraphs contained between 50 and 200 words.

Recall	Precision (27 queries)				
	Document	Bounded		Combined	
	(D)	Paragraph (BP)		(2∗D+BP)	
0	28.2	26.6	(−5.9)	28.5	(+0.9)
10	28.2	26.6	(−5.9)	28.5	(+0.9)
20	28.2	26.6	(−5.9)	27.9	(-1.3)
30	23.6	21.9	(−7.2)	23.8	(+1.1)
40	21.4	20.3	(−5.3)	21.4	(+0.1)
50	21.4	19.8	(−7.5)	20.3	(-4.9)
60	13.1	15.7	(+19.7)	12.7	(-3.6)
70	10.3	10.2	(−0.6)	11.1	(+8.2)
80	9.6	9.6	(+0.1)	10.2	(+6.0)
90	8.9	9.2	(+3.8)	9.6	(+8.1)
100	7.5	8.2	(+9.1)	8.5	(+12.6)
avg	18.2	17.7	(-2.9)	18.4	(+1.0)

Table 2. TIPSTER volumes 1 and 2, queries 101-150, TIPSTER 24 month judgements. Paragraphs contained between 50 and 200 words.

Interpolated Recall - Precision Averages:					
	Document	Bounded		Combined	
	(D)	Paragraph (BP)		(2∗D+BP)	
at 0.00	0.8275	0.7067	(−14.6)	0.8400	(+1.5)
at 0.10	0.5945	0.4722	(−20.6)	0.6101	(+2.6)
at 0.20	0.5383	0.4218	(−21.6)	0.5512	(+2.4)
at 0.30	0.4775	0.3673	(−23.1)	0.4985	(+4.4)
at 0.40	0.4290	0.3170	(−26.1)	0.4468	(+4.2)
at 0.50	0.3720	0.2659	(−28.5)	0.3864	(+3.9)
at 0.60	0.3133	0.2055	(−34.4)	0.3265	(+4.2)
at 0.70	0.2316	0.1393	(−39.9)	0.2482	(+7.2)
at 0.80	0.1388	0.0758	(−45.4)	0.1652	(+19.0)
at 0.90	0.0588	0.0293	(−50.2)	0.0758	(+28.9)
at 1.00	0.0095	0.0046	(−51.2)	0.0082	(-13.7)
Average precision (non-interpolated) over all rel docs					
	0.3480	0.2511	(−27.9)	0.3629	(+4.3)

documents (1987 Wall Street Journal), long documents (65 megabytes of Federal Register), and mixed documents (a combination of 1987 Wall Street Journal and 65 megabytes of Federal Register).

Manual verification of the heuristics revealed large variation in the sizes of paragraphs in the Wall Street Journal and the Federal Register. Wall Street Journal paragraphs were sometimes just one sentence long, and didn't always correspond to a shift in content. Federal Register paragraphs ranged from one line to thousands of words, and indicated more consistently a shift in content.

2.2 Bounded-Paragraph Passages

Bounded paragraphs were an attempt to solve the problems described above. The hypothesis was that one can provide more consistent paragraphs by merging short paragraphs and by dividing large paragraphs. A series of experiments suggested that 50 words be a minimum paragraph size and 200 words be a maximum paragraph size. Each paragraph smaller than the minimum was merged into the paragraph that followed it. In preliminary experiments with the 1987 Wall Street Journal, Federal Register, and combined collections described above, bounded paragraphs were more effective than "real" paragraphs.

This approach was evaluated in an experiment with 65 megabytes of TIPSTER volume 1 Federal Register documents (Table 1), and on TIPSTER volumes 1 and 2 as as part of the UMass participation in the TIPSTER project (Table 2).[1] In these experiments, retrieval of documents based on their single best

[1] TIPSTER participants were evaluated on only the top 1000 documents submitted for each query, so the results shown are not directly comparable to the full-ranking results from other tables in this paper.

matching paragraph produced losses of 2.9% and 27.9% respectively. However, combining the document-level and paragraph-level evidence produced 1.0% and 4.3% improvements in average precision.

The results from the Federal Register experiment should be viewed with caution, because there were very few relevant documents for each query. About half of the queries had just one relevant document. The results are included here because they are counterintuitive. One might expect queries with so few relevant documents to be unstable, but they were not. Instead, passage-level evidence had relatively little impact.

It is unclear from these experiments why retrieval based upon the best bounded paragraph performed so poorly with TIPSTER volumes 1 and 2. Another experiment with the 1987 Wall Street Journal and 65 megabytes of Federal Register (not included here) produced very similar results. One possibility is that the length-based criteria used for merging and dividing paragraphs produced passages that did not sufficiently organize text by content. An approach to reconstructing paragraphs based on semantic considerations (e.g. TextTiling [9]) might yield better results.

2.3 Window Passages

Discourse and semantic passages are based on the assumption that there is a single "good" organization of the information in a document. The difference between the two approaches is whether the writer is trusted to reveal that organization. However, what if there is no single division of a document into passages that is appropriate for all queries?

Passages based on fixed-length windows appear initially to have even more problems, because they impose a single division of text into passages, and because they are more likely to divide relevant text among different passages. However, both of these problems are solved easily.

The solution adopted for INQUERY was to begin the first passage in a document at the first term matching the query, creating new passages of length n every $\frac{n}{2}$ words. For example, if the passage length was 100 and the first matching term was at position 62, overlapping passages of 100 words would begin at positions 62, 112, 162, 212, etc. The use of overlapping passages reduces the chance that a small block of relevant text is split among two passages; it is similar in spirit to passage *blurring* [15].

If passage locations vary from query to query, then the formulae used to determine the evidence contributed by a term t to a document d cannot be used directly for passages. These formulae, shown below, require the frequency of the most frequent term in each document (max_tf_d).

$$ntf_{td} = 0.4 + 0.6 * \frac{\log(tf_{td} + 0.5)}{\log(max\_tf_d + 1.0)} \quad (1)$$

$$idf_t = \frac{\log\left(\frac{\text{docs in collection}}{\text{docs containing } t}\right)}{\log(\text{docs in collection})} \quad (2)$$

$$belief_{td} = 0.4 + 0.6 * ntf_{td} * idf_t \quad (3)$$

max_tf is determined easily during indexing for each document, but cannot be determined during indexing for passages whose locations depend upon queries. It is impractical to find the max_tf of a passage at runtime, so an alternative is required. The simplest solution is to use the max_tf of the document. However, better results are obtained when the term frequency of the passage (tf_{tp}) is scaled by the passage length. It is unclear why passage length, which is essentially constant for fixed-length windows, should be preferable for ranking passages.

The window-based approach to passages was tested in experiments with four collections. The collections, and their characteristics, are shown below.

Name	Size	Documents	Queries	Avg Query Length	Results (Table)
TIPSTER Fed Reg	474 MB	46,315	38	42.4	3
West	298 MB	11,953	34	11.3	4
TIPSTER vol. 1	1.2 GB	510,887	50	42.7	5
NPL	3 MB	11,429	93	10.8	6

The Federal Register experiment described in Section 2.2 found that passages based on bounded paragraphs had little effect on document retrieval (Table 1). Prior to testing the effects of window passages, a possible flaw in that earlier experiment, the small number of relevance judgements for several queries, was corrected. A new set of 38 queries was assembled, in which each query had at least 5 relevant documents. This is a better methodology, but unfortunately makes the results obtained not directly comparable to those in Table 1.

Table 3 shows the results from the experiment with window passages and the revised set of Federal Register queries. This experiment found that document retrieval based upon evidence from the single best passage was 20.7% better than retrieval based upon evidence from the entire document. A weighted

Table 3. TIPSTER Federal Register volumes 1 and 2, selected queries, TREC-1 and TREC-2 judgements. The window size was 300 words.

Recall	Precision (38 queries)				
	Document	Window		Combined	
	(D)	(W)		(D+7*W)	
0	66.4	68.1	(+2.6)	72.5	(+9.1)
10	53.1	60.1	(+13.1)	59.9	(+12.7)
20	40.4	48.9	(+21.0)	50.3	(+24.4)
30	36.2	43.5	(+20.2)	44.8	(+23.8)
40	32.5	40.1	(+23.6)	41.2	(+26.9)
50	26.7	36.1	(+35.1)	36.2	(+35.8)
60	22.7	33.6	(+47.7)	33.0	(+45.4)
70	20.2	28.1	(+39.2)	28.5	(+41.6)
80	18.4	23.2	(+25.9)	24.8	(+34.4)
90	13.4	16.4	(+21.9)	16.2	(+20.6)
100	8.9	11.2	(+25.5)	11.1	(+24.6)
avg	30.8	37.2	(+20.7)	38.0	(+23.5)

Table 4. West collection. The passage size was 200 words.

Recall	Precision (50 queries)				
	Document	Window		Combined	
	(D)	(W)		(D+2*W)	
0	87.6	95.9	(+9.5)	93.0	(+6.1)
10	81.7	86.6	(+6.0)	87.9	(+7.5)
20	77.1	79.3	(+2.9)	79.8	(+3.5)
30	72.4	74.9	(+3.4)	76.5	(+5.7)
40	64.1	65.2	(+1.8)	66.9	(+4.4)
50	59.3	58.2	(–1.9)	61.1	(+3.0)
60	52.0	51.3	(–1.4)	54.1	(+4.0)
70	43.5	43.9	(+0.9)	45.1	(+3.8)
80	35.2	37.5	(+6.7)	38.0	(+8.0)
90	21.5	23.7	(+10.2)	24.3	(+12.7)
100	12.0	12.9	(+7.5)	13.1	(+9.2)
avg	55.1	57.2	(+3.8)	58.2	(+5.5)

combination of document-level and passage-level evidence (D+7*W) yielded a 23.5% improvement. Significant improvements were observed at virtually all recall levels. The heavy emphasis on passage-level evidence was required due to the large disparity in the performance of document-level and passage-level evidence. If the weighting were more moderate, for example D+2*W, the combination would have been slightly worse than passage-level evidence alone.

A second experiment investigated the effect of window passages on a collection of legal documents provided by the West Publishing Company. The West documents are almost four times longer than the Federal Register documents, the queries are almost four times shorter, and the average precision from document-level evidence relatively high (55.1%).

The results were that document retrieval based upon evidence from the single best passage was 3.8% better than retrieval based upon evidence from the entire document (Table 4). A weighted combination of document-level and passage-level evidence (D+2*W) yielded a 5.5% improvement. Once again, significant improvements were seen at most recall levels.

A third experiment investigated the effect of window passages on TIPSTER volume 1, a heterogeneous collection of about 500,000 documents. The TIPSTER documents range from 1 sentence to thousands of words in length, and contain a varied vocabulary. The earlier experiment with TIPSTER volumes 1 and 2, described in Section 2.2, found passages based upon bounded paragraph to be detrimental by themselves, and somewhat helpful when combined with document-level evidence (Table 2). The experiment with window passages used a different set of queries, a different portion of the TIPSTER collection, and a different method of evaluation, so the results of the two experiments are indicative but not directly comparable.

Table 5 shows the results from the experiment with window passages on TIPSTER volume 1. This experiment found that document retrieval based upon evidence from the single best passage was about the same as retrieval based upon evidence from the entire document. Combining document-level and

Table 5. TIPSTER volume 1, queries 51-100, TREC-1 judgements. The passage size was 200 words.

Recall	Precision (50 queries)				
	Document	Window		Combined	
	(D)	(W)		(D+2*W)	
0	82.7	86.6	(+4.7)	86.5	(+4.5)
10	60.3	62.5	(+3.6)	65.0	(+7.8)
20	52.6	51.1	(−2.9)	55.9	(+6.2)
30	46.7	44.4	(−4.9)	49.4	(+5.9)
40	40.4	39.7	(−1.9)	43.6	(+7.8)
50	34.8	34.6	(−0.7)	37.4	(+7.5)
60	30.4	30.0	(−1.3)	32.8	(+8.0)
70	25.4	24.6	(−3.3)	27.7	(+9.1)
80	19.8	18.5	(−6.3)	21.4	(+8.6)
90	12.1	12.4	(+2.6)	13.9	(+14.6)
100	2.4	2.9	(+17.8)	2.9	(+21.0)
avg	37.1	37.0	(−0.1)	39.7	(+7.1)

Table 6. NPL collection. The passage size was 50 words.

Recall	Precision (93 queries)				
	Document	Window		Combined	
	(D)	(W)		(D+2*W)	
0	71.1	72.3	(+1.7)	73.9	(+3.9)
10	57.5	61.1	(+6.3)	62.2	(+8.3)
20	47.5	48.8	(+2.8)	50.8	(+6.9)
30	39.0	39.7	(+1.6)	41.7	(+6.8)
40	32.4	32.5	(+0.3)	34.7	(+7.2)
50	25.2	26.4	(+4.9)	27.1	(+7.8)
60	18.1	18.7	(+3.4)	19.6	(+8.2)
70	13.4	14.6	(+9.0)	14.8	(+10.5)
80	10.5	10.9	(+4.6)	11.2	(+6.9)
90	6.3	6.5	(+3.2)	6.6	(+4.9)
100	3.2	3.2	(−1.9)	3.2	(−1.0)
avg	29.5	30.4	(+3.3)	31.4	(+6.7)

passage-level evidence yielded a 7.1% improvement. Significant improvements were seen at almost every level of recall for the combination retrieval.

The fourth and final experiment investigated the effect of long window passages on a collection of very short documents. The goal was to determine whether use of window passages, with the slightly different method of calculating belief, had a negative effect on short documents. The NPL collection of short physics abstracts was used for this fourth experiment.

One result was that the passage-level and document-level approaches to calculating belief yielded similar results, although small differences were apparent at all recall levels (Table 6). However, combining the results of the two approaches yielded a 6.5% increase in average precision, with significant improvements occurring at most recall levels. This result was unexpected, and bears further study. It may suggest ways of improving INQUERY's method of combining evidence, or it may mean that considering the same evidence from two points of view, even if they are only slightly different, yields some advantage.

2.3.1 Window Size

The four experiments described above were conducted with window sizes of 300 for Federal Register, 200 for West, 200 for TIPSTER volume 1, and 50 for NPL. These sizes were chosen to provide the best results for each collection. However, they raise the question of whether there is a single "right" passage size.

Experiments were conducted with a wide variety of passage sizes for each collection, ranging from 25 words to 10,000 words. Tables 7 and 8 show some of the results. Passages of 150-300 words yielded the best results. Performance was reasonably consistent in that range. If one had to choose a single size for all collections, a window of 200 or 250 words provides consistently good results.

Two experiments were conducted to determine the effects of combining the results of document-level evidence with evidence from 2 passages of different sizes. Both experiments yielded further improvements of about 1%, over all recall levels and in average precision, on the West collection. One can view this

Table 7. Effect of passage size on document retrieval, using only passage-level evidence. Values shown are 11-point average precision.

	Size of Passage											
Collection	25	50	75	100	150	200	250	300	400	500	750	1000
FR	32.0	33.9	33.4	34.8	36.0	36.0	36.6	37.2	36.5	35.7	35.2	34.9
West	51.6	54.1	55.6	56.0	56.5	57.2	57.3	56.5	56.6	56.2	55.9	55.0
Tip1	30.0	32.1	33.9	35.2	36.2	37.0	37.6	37.6	37.6	37.2	37.0	36.1
NPL	29.6	30.4	29.8	29.7	29.6	29.6	29.6	29.7	29.8	29.7	29.8	29.8

Table 8. Effect of passage size on document retrieval, using a combination of document and passage-level evidence (D+2*P). Values shown are 11-point average precision.

	Size of Passage											
Collection	25	50	75	100	150	200	250	300	400	500	750	1000
FR	35.2	36.2	35.6	35.8	36.6	36.8	36.8	36.9	35.9	35.6	35.2	34.8
West	57.3	57.8	57.7	57.9	58.0	58.2	58.0	57.6	57.5	57.4	56.7	56.0
Tip1	39.0	39.3	39.7	40.0	40.0	39.7	39.6	39.6	39.3	38.8	38.3	37.8
NPL	31.2	31.4	31.2	31.1	31.0	31.0	31.1	31.1	31.1	31.1	31.2	31.2

result as either diminishing returns or an indication that further improvements are possible. With careful implementation, this improvement might come at little additional cost.

3 Implementation and Efficiency

The original implementation of INQUERY repeatedly read from disk the complete inverted record for a term, merged the evidence into its accumulating store of evidence for each document, and then discarded the inverted list. This approach minimizes I/O, but requires increasing amounts of memory as a collection grows. Treating passages like documents would require at least an order of magnitude more memory.[2]

INQUERY now reads inverted lists in blocks, in parallel, computing a complete score for one document before proceeding to the next. This approach is known to slow I/O [18], but scales well, and enables optimizations that are impossible with the previous "term at a time" approach. Several commercial IR vendors also use this technique.[3]

The window-based approach to passages described above requires that a document be divided into passages after the query is available. Many document passages contain few or no query terms, so it makes little sense to evaluate the entire query once per passage. Instead, INQUERY assigns a default belief to each passage, and then each query operator updates the belief of any passage for which it has evidence. The result is a relatively fast ability to rank every passage in a document. The cost of ranking passages for 50 queries on TIPSTER volume 1 is about 25% higher than the cost of ranking documents.

One can avoid ranking every passage by ranking every document and then ranking passages for only those documents that exceed some threshold [14]. However, this technique runs the risk of missing a highly relevant passage in a long, mostly irrelevant document. One might expect this problem to become more serious as document length increases.

The results of experiments with the Federal Register collection, described above (Table 3), were examined to determine whether ranking passages only in the top-ranked documents would cause the system to overlook some relevant documents. The hypothesis was that if the system were required to retrieve n documents, it would be better to rank all documents by their best passages and then return the top n documents than to rank all documents and then rerank the top n based on their best passages. This collection was chosen because the documents are long and because passage-level evidence yielded significantly better results than document-level evidence (Table 3). There were 38 queries, and 1085 relevant documents (28.6 relevant documents per query, on average).

If the threshold n is 250 documents, the approach based on ranking all passages finds 33 more relevant documents. If the threshold is 500, ranking all passages finds 28 more relevant documents. The difference drops to 7 with a threshold of 1000 documents, and 4 with a threshold of 2000 documents. This result suggests that ranking all documents and then reranking the top n by passages may cause relatively little loss in recall, as long as n is large relative to the average number of relevant documents.

[2] Assuming 10 passages per document, which is conservative for the TIPSTER and West collections.

[3] Personal communications.

4 Conclusions

This paper describes experiments with paragraph-based and window-based methods of defining passages. The experiments were conducted with three relatively homogeneous collections (Federal Register, West, and NPL) and two relatively heterogeneous collections (TIPSTER volume 1, and TIPSTER volumes 1 and 2).

The experiments were intended to help understand how passages should be defined, how they should be ranked, and how they should be used in retrieval. Although the experiments raise new questions, some general conclusions can be drawn from the results reported here.

Passages based upon paragraph boundaries were less effective than passages based upon overlapping text windows of varying sizes. This result held for both document retrieval based on a single best passage, and document retrieval based upon combining document-level and passage-level evidence. This result may be due to inconsistent use of paragraphs by the authors of TIPSTER documents, or it may be that there is no single division of documents into passages that is appropriate for all queries. It may also be that paragraphs are too small, and that a larger unit of discourse (e.g. section or subsection) would yield better results.

A slight variation of the formulae for ranking documents was effective for ranking passages. However, it is not clear why scaling the term frequency in the passage (tf_p) by a constant value should be more effective than scaling it by the maximum term frequency in the document (max_tf_d). Improvements may also be possible by finding an alternative to idf_t that reflects inverse passage frequency. Finally, the improvement seen in the NPL experiment suggests that further improvements may also be possible in the formulae for ranking documents.

Perhaps the clearest result of the experiments is that it was always best to combine document-level evidence and passage-level evidence. This result confirms some of the previous research on passages, and also confirms that the effectiveness of the inference network generally improves with the amount of evidence available. The experiments leave open the question of how best to weight the different sources of evidence, and whether it is appropriate to conduct retrieval by gathering and combining evidence from several levels of a document's structure.

Acknowledgements

I thank the West Publishing Company for providing the West collection of legal texts. This research was supported by the NSF Center for Intelligent Information Retrieval at the University of Massachusetts.

References

1. N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. The effect of multiple query representations on information retrieval system performance. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 339–346, Pittsburgh, PA, June 1993. Association for Computing Machinery.

2. C. Buckley, J. Allan, and G. Salton. Automatic routing and ad-hoc retrieval using SMART: TREC-2. In D. Harman, editor, *Proceedings of the Second Text REtrieval Conference (TREC-2)*. National Institute of Standards and Technology Special Publication 500-215, 1994.

3. J. P. Callan and W. B. Croft. An evaluation of query processing strategies using the TIPSTER collection. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 347–356, Pittsburgh, PA, June 1993. Association for Computing Machinery.

4. J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83, Valencia, Spain, 1992. Springer-Verlag.

5. W. B. Croft, J. Callan, and J. Broglio. TREC-2 routing and ad-hoc retrieval evaluation using the INQUERY system. In D. Harman, editor, *Proceedings of the Second Text REtrieval Conference (TREC-2)*. National Institute of Standards and Technology Special Publication 500-215, 1994.

6. Cary Griffith. WESTLAW's winning ways. *Law Office Computing*, pages 31–38, February/March 1993.

7. David Haines and W. B. Croft. Relevance feedback and inference networks. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–11, Pittsburgh, PA, June 1993. Association for Computing Machinery.

8. D. Harman. The DARPA Tipster project. *SIGIR Forum*, 26(2):26–28, 1992.

9. M. A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 59–68, Pittsburgh, PA, June 1993. Association for Computing Machinery.

10. A. Moffat, R. Sacks-Davis, R. Wilkinson, and J. Zobel. Retrieval of partial documents. In D. Harman, editor, *Proceedings of the Second Text REtrieval Conference (TREC-2).* National Institute of Standards and Technology Special Publication 500-215, 1994.

11. J. O'Connor. Answer-passage retrieval by text searching. *Journal of the American Society for Information Science*, 31(4):227–239, 1980.

12. J. S. Ro. An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval. I. On the effectiveness of full-text retrieval. *Journal of the American Society for Information Science*, 39(2):73–78, 1988.

13. G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA, June 1993. Association for Computing Machinery.

14. G. Salton and C. Buckley. Automatic text structuring and retrieval – Experiments in automatic enclopedia searching. In A. Bookstein, Y. Chiaramella, G. Salton, and V. V. Raghavan, editors, *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Chicago, IL, October 1991. Association for Computing Machinery.

15. C. Stanfill and D. L. Waltz. Statistical methods, Artificial Intelligence, and Information Retrieval. In P. S. Jacobs, editor, *Text-based intelligent systems*, pages 215–225. Lawrence Erlbaum, 1992.

16. H. R. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, 1991.

17. R. Wilkinson. Effective retrieval of structured documents. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, July 1994. Association for Computing Machinery.

18. P. Willett. A nearest neighbor search algorithm for bibliographic retrieval from multilist files. *Information Technology*, 3(2):78–83, 1984.