Incorporating Window-Based Passage-Level Evidence in Document Retrieval

Wensi Xi, Richard Xu-Rong, Christopher S.G. Khoo

Center for Advanced Information Systems
School of Applied Science
Nanyang Technological University
Singapore, 639798

Keywords: passage retrieval, text segmentation, merging search results/information synthesis

Correspondence should be addressed to:

Chris Khoo
School of Applied Science
Nanyang Technological University
Blk N4 #2A-32, Nanyang Ave
Singapore 639798
Republic of Singapore

Email: assgkhoo@ntu.edu.sg
Tel: (65) 7904602
Fax: (65) 7926559

Incorporating Window-Based Passage-Level Evidence in Document Retrieval

Abstract. This study investigated whether information retrieval can be improved if documents are divided into smaller subdocuments or passages, and the retrieval score for these passages are incorporated in the final retrieval score for the whole document. The documents were segmented by sliding a window of a certain size across the document. Each time the window stopped, it displayed/extracted a certain number of contiguous words. A retrieval score was calculated for each of the passages extracted, and the highest score obtained by a passage of that size was taken as the document's "window score" for that window size. A range of window sizes were tried.

The experimental results indicated that using a fixed window size of 50 gave better results than other window sizes for the TREC test collection. This window size yielded a significant retrieval improvement of 24% compared to using the whole-document retrieval score. However, combining this window score and the whole-document retrieval score did not yield a retrieval improvement.

Identifying the highest window score for each document (using window sizes varying from 50 to 400 words), and adopting it as the document retrieval score yielded a retrieval improvement of about 5% over taking the size-50 window score. Different window sizes were found to work best for different queries. If we could predict accurately the best window size to use for each query, a maximum retrieval improvement of 42% could be obtained. However, an effective way has not been found for predicting which window size would give the best results for each query.

Keywords: passage retrieval, text segmentation, merging search results/information synthesis

1 Introduction

Large collections of full-text documents are now commonly used in automated information retrieval. When the stored documents are long, only a small section of the document may be relevant to the user, and it may be better to present this section to the user. With whole-document retrieval, the high relevance of a small section may be obscured by the irrelevance of the rest of the document and may not be reflected adequately in the whole document retrieval score. Furthermore, query terms occurring within a small document passage are more likely to be in the desired syntagmatic relations with one another as required by the query, than if the words are distributed across a long document. This suggests that it is desirable to identify smaller passages in a document that are highly likely to be relevant to the user. The passage-level retrieval score can be combined with the whole-document retrieval score to better predict the likelihood that the document contains relevant information.

This study investigated whether information retrieval can be improved if documents are divided into smaller subdocuments or passages, and the retrieval score for these passages incorporated in the final retrieval score for the whole document. The approach taken was to segment the passages by extracting a fixed number of contiguous words from the whole document. The process can be viewed as using a window of a certain size (i.e. number of words) that slides through the whole document. Each time the window stops, it displays or extracts a certain number of contiguous words in the document. A retrieval score was calculated for each of the passages extracted, and the highest score obtained by a passage of that size was taken as the document's "window score" for that window size. A range of window sizes were tried. For a particular query, a set of retrieval scores were calculated for each document: the whole-document retrieval score and one score for each window size. The scores were then combined in some way to give the composite retrieval score for the document.

We investigated the following methods for selecting a window score as the document retrieval score:

1 Using a fixed window size for all queries. We investigated a range of window sizes and identified the size that gave the best retrieval results.

2 Using the highest window score for each document. The window size that obtains the highest retrieval score is identified for each document, and the retrieval score for this window size is considered the highest window score for the document. The purpose is to identify the passage (regardless of size) in the document with the highest retrieval score.

3 Using the best window size for each query. The window size is fixed for a query, but may vary from query to query.

We also attempted to combine the various passage-level scores with the whole-document score to obtain a composite retrieval score.

2. Previous Studies

Other researchers have developed methods for passage retrieval and studied how passage-level evidence can be incorporated in document retrieval. These studies have used different approaches to segmenting a document into passages.

Moffat, Sack-Davis, Wilkinson & Zobel (1994, 1995) suggested that passages be obtained by partitioning documents into disjoint segments of roughly equal length. In these experiments, the passages were generated by gathering adjacent paragraphs until each agglomeration was at least some fixed number of bytes. Thus a page boundary always coincided with a paragraph boundary. Their experiments showed that retrieval based on passages with minimum length in the range 1000-2000 bytes (or roughly 150-300 words) was significantly more effective than whole-document retrieval. These results were based on the Federal Register (FR) subcollection of TREC, which contains many long documents. They further experimented with structured documents and found that the techniques they had developed for unstructured text could be used for structured documents almost without any change.

Hearst and Plaunt (1993) treated full-length documents as composed of a sequence of locally concentrated discussions. They suggested that passages consist of sequences of sentences, where the boundaries between passages or tiles were determined by automatically-detected shift of topic. Their strategy was to divide the documents into motivated segments, retrieve the top-scoring 200 segments that most closely match the query according to the vector space model, and then sum the vectors for all segments that are from the same document. This causes the parts of the documents that are most similar to the queries to contribute to the final retrieval score for the document. This approach was found to work significantly better than either whole-document retrieval or single segment retrieval.

Mittendorf and Schauble (1994, 1995) also suggested using inferred passage boundaries, by employing a hidden Markov model to determine passages appropriate to each query. They found that passage ranking improved retrieval effectiveness.

Other researchers have used simpler methods for determining passage boundaries. Salton, Allan & Buckley (1993) used document features such as paragraphs and section boundaries for passage retrieval. They found that individual sentences can help determine the relevance of the whole document. They suggested using the document markup tags to determine passages (sections, paragraphs, sentences) in a two-pass method. First the overall or global similarity between the documents and query are calculated to obtain the whole-document retrieval score. Documents with scores exceeding a threshold are shortlisted and passage-level evidence (local similarity scores) is used to refine the ordering. The shortlisted documents that contain a sufficient number of passages that are similar to the query are assumed to be relevant. They found that this approach improved retrieval effectiveness, although it may exclude long documents with only a small block of relevant material.

Wilkinson (1994) also suggested that document markup or logical structure be used to delimit passages or sections, and explored both the ability of sections to select relevant documents and the use of similarity functions to select relevant sections. Although the retrieval results of using structure as a basis for whole document retrieval was mixed, the results indicated some advantages of passage retrieval, e.g. sections did provide a valuable mechanism for identifying the interesting parts of long documents.

Callan (1994) found that passages based on paragraph boundaries were less effective than passages based on overlapping fixed-length windows. This approach, which eliminates the problems of relying on semantic or structural features of documents (e.g. paragraphs or sentences), is the approach adopted in our study. Callan’s work was carried out using the INQUERY system, a probabilistic information retrieval system. His experimental results suggest that different window sizes worked best for different document collections. He said that, generally, passages of 150-300 words yielded the best results. Using window passages of a fixed-size gave better results than whole-document retrieval for 3 homogeneous document collections that he used, but not for the TREC 1 and 2 test collections which includes several corpora. However, a linear combination of passage-level and whole-document scores consistently gave the best results.

3. Research Method

This study uses the TREC-5 and TREC-6 test collections (URL http://trec.nist.gov/), comprising 100 queries (topic 251-350) and 8 corpora – Financial Times, Wall Street Journal, Associated Press Newswire, Congressional Record, Federal Register, ZIFF (materials from the Ziff Communications Company), Foreign Broadcast Information Service, and Los Angeles Times. The test collection has altogether 634,939 documents, of which 10,135 documents are relevant to one or more queries.

For each retrieval run, two versions of TREC queries were used: long queries and short queries. A TREC query, as the example in Table 1 illustrates, has three fields: *title*, *description,* and *narrative*. For long queries, we used the text in all 3 fields. For short queries, we included only the *title* and *description* fields.

Table 1. An example of a TREC query

Number:	251
Title:	Exportation of Industry
Description:	Documents will report the exportation of some part of U.S. Industry to another country.
Narrative:	Relevant documents will identify the type of industry being exported, the country to which it is exported; and as well will reveal the number of jobs lost as a result of that exportation.

The experiments make use of an in-house developed retrieval system based on the vector space model. The *ntf**idf* weighting scheme with cosine normalization was used for constructing the query vectors and the *tf**idf* weighting scheme with cosine normalization was used for the document vectors (Harman, 1992; Salton & Buckley, 1988). *tf* refers to the term frequency (the number of times the term occurs in the document or query), and *idf* refers to the inverse of the document frequency (the number of documents in the database containing the term). *ntf* refers to *normalized term frequency* and is given by the formula

$$ntf = 0.5 + 0.5 * tf / max\_tf$$

where *max_tf* is the highest term frequency obtained by terms in the query. The retrieval score for the whole document is calculated by taking the inner product of the document and query vectors.

For each query, this whole-document retrieval score was used to filter out the top-ranked 2000 documents. Only these top 2000 documents retrieved for each query were used in the experiments with passage retrieval. However, the weighting described above was also used to calculate the passage-level retrieval scores.

In the experiments with passage retrieval, each document was segmented into overlapping passages of a fixed number of words by “sliding” a window of a particular size across the document. We shall refer to these passages as “window passages”. Adjacent window passages overlap by 50%, i.e. the middle of the current window becomes the origin of the next window. For each passage extracted from the document, a retrieval score was calculated for the passage using the same method as that used for calculating the whole-document retrieval score. A vector was thus constructed to represent the passage, and its inner product similarity with the query vector was calculated. For a particular window size, the window passage with the highest retrieval score in the document was taken as the document’s passage-level score for that window size.

Window sizes of 50, 100, 150, 200, 250, 300, 400, 500 and 800 words were used in the study. For each document, a retrieval score was calculated for each window size. Three experiments were carried out:

- *Experiment 1*: using a fixed window size for all queries. We investigated whether there was a particular window size that in general gives the best results and that outperforms whole document retrieval.
- *Experiment 2*: using the highest window score for each document. We investigated whether identifying the window size with the highest retrieval score for a document and adopting that score as the document retrieval score gives better results than using a fixed

window size as in Experiment 1. Before selecting the best window score, the scores were first normalized. *Log_2* was applied to all the scores. The scores were then converted to z-scores using the formula: $(X_{i,win} – M_{win}) / \sigma_{win}$
where $X_{i,win}$ is the passage score obtained by document *i* for window size *win*, M_{win} is the mean passage score obtained by all the 2000*100 documents (the top 2000 documents filtered out for each query) for window size *win*, and σ_{win} is the standard deviation of the scores for window size *win*.

- *Experiment 3*: using the best window size for each query, i.e. using a fixed window size for each query, but the size may be different for different queries. We investigated whether different window sizes work best for different queries, and whether the best window size for a query can be identified automatically.

4. Results

4.1 Experiment 1: using a fixed window size for all queries

Table 1 shows the retrieval results for long queries – for whole-document retrieval as well as for using the retrieval scores for window sizes ranging from 50 words to 800 words. Table 2 gives the retrieval results for short queries. The window size of 50 gave the best results. The improvement in the non-interpolated average precision compared with whole-document retrieval was 24.2% for long queries and 22.0% for short queries. The improvement for both long and short queries was significant at the 0.05 level (using a 2-tailed t-test). The recall-precision curves for window size 50 and whole-document retrieval are shown in Fig. 1 (for long queries) and Fig. 2 (for short queries). It can be seen that there was an improvement throughout the recall-precision curve.

The results shown in Tables 1 and 2 also indicate that for all the window sizes, passage retrieval outperformed whole-document retrieval. It is also observed that for short queries, the window size of 150 gave the best precision at 0% and 0.1% recall levels, i.e. at the top-ranked documents.

Callan (1994) didn’t find that window passage scores improved retrieval compared with whole-document retrieval, using the TREC 1 and 2 test collection. However, he found that combining a window score with the whole document retrieval score yielded a 7% retrieval improvement. However, we didn’t manage to get a retrieval improvement by taking a linear combination of a window score and the whole-document score.

Table 1. Recall-precision figures for 100 long queries

	Recall	**Whole document retrieval**	**W50**	**W100**	**W150**	**W200**	**W250**	**W300**
Interpolated Precision for 11 Recall Points	0.00	0.5738	0.5942	0.5829	0.5592	0.5824	0.5800	0.5806
	0.10	0.3241	0.3407	0.3577	0.3524	0.3505	0.3427	0.3471
	0.20	0.2345	0.2716	0.2580	0.2643	0.2570	0.2615	0.2559
	0.30	0.1477	0.2091	0.1971	0.1891	0.1843	0.1843	0.1807
	0.40	0.1158	0.1586	0.1496	0.1365	0.1372	0.1317	0.1369
	0.50	0.0717	0.1039	0.1050	0.0960	0.0941	0.0903	0.1012
	0.60	0.0344	0.0639	0.0642	0.0526	0.0492	0.0469	0.0558
	0.70	0.0153	0.0388	0.0341	0.0232	0.0224	0.0204	0.0305
	0.80	0.0092	0.0188	0.0182	0.0129	0.0100	0.0089	0.0186
	0.90	0.0075	0.0129	0.0125	0.0084	0.0057	0.0063	0.0161
	1.00	0.0075	0.0129	0.0125	0.0084	0.0057	0.0063	0.0161
Non-interpolated Average Precision		0.1161	0.1442	0.1404	0.1345	0.1319	0.1314	0.1371

	Recall	**W400**	**W500**	**W800**	**Highest Window Score**	**Predicted Best Window Size for Query**	**Best Window Size for Query**	
Interpolated Precision for 11 Recall Points	0.00	0.5929	0.5860	0.5928	0.5873	0.6196	0.6976	
	0.10	0.3498	0.3586	0.3555	0.3654	0.3724	0.4014	
	0.20	0.2518	0.2561	0.2497	0.2886	0.2905	0.3065	
	0.30	0.1640	0.1671	0.1660	0.2162	0.2087	0.2259	
	0.40	0.1228	0.1253	0.1268	0.1568	0.1603	0.1762	
	0.50	0.0849	0.0914	0.0866	0.1164	0.1064	0.1179	
	0.60	0.0424	0.0435	0.0449	0.0706	0.0638	0.0722	
	0.70	0.0187	0.0185	0.0191	0.0390	0.0388	0.0407	
	0.80	0.0086	0.0082	0.0095	0.0204	0.0188	0.0187	
	0.90	0.0063	0.0059	0.0071	0.0158	0.0165	0.0166	
	1.00	0.0063	0.0059	0.0071	0.0158	0.0165	0.0166	
Non-interpolated Average Precision		0.1271	0.1288	0.1274	0.1511	0.1520	0.1645	

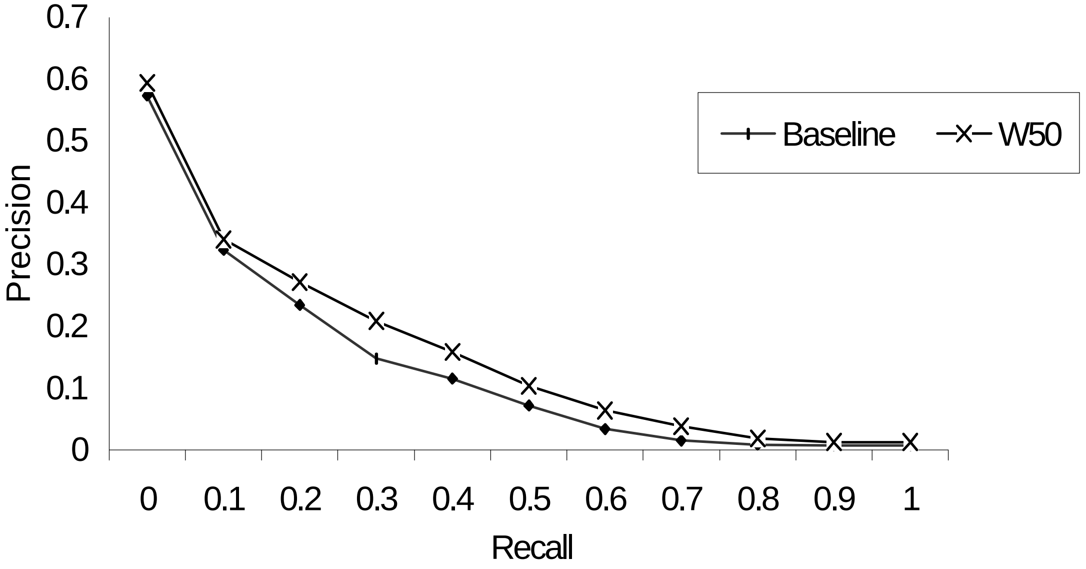

Figure 1. Recall-precision graph for long queries

Table 2. Recall-precision figures for 100 short queries

	Recall	**Whole document retrieval**	**W50**	**W100**	**W150**	**W200**	**W250**	**W300**
Interpolated Precision for 11 Recall Points	0.00	0.5077	0.5094	0.5152	0.5464	0.5275	0.5482	0.5298
	0.10	0.2634	0.2816	0.2919	0.3058	0.3106	0.2873	0.2815
	0.20	0.1784	0.2274	0.2147	0.2153	0.2098	0.2155	0.1946
	0.30	0.1320	0.1755	0.1601	0.1578	0.1511	0.1541	0.1422
	0.40	0.0970	0.1440	0.1290	0.1136	0.1153	0.1089	0.1018
	0.50	0.0618	0.0975	0.0852	0.0827	0.0810	0.0708	0.0690
	0.60	0.0346	0.0617	0.0528	0.0458	0.0486	0.0448	0.0420
	0.70	0.0187	0.0396	0.0328	0.0255	0.0291	0.0264	0.0248
	0.80	0.0096	0.0105	0.0116	0.0096	0.0109	0.0108	0.0113
	0.90	0.0091	0.0038	0.0034	0.0058	0.0073	0.0092	0.0091
	1.00	0.0091	0.0038	0.0034	0.0058	0.0073	0.0092	0.0091
Non-interpolated Average Precision		0.0983	0.1199	0.1139	0.1137	0.1138	0.1108	0.1056
	Recall	**W400**	**W500**	**W800**	**Highest Window Score**	**Predicted Best Window Size for Query**	**Best Window Size for Query**	
Interpolated Precision for 11 Recall Points	0.00	0.5089	0.5224	0.5217	0.5542	0.5425	0.6576	
	0.10	0.2764	0.2854	0.2804	0.3036	0.3212	0.3389	
	0.20	0.1898	0.1848	0.1804	0.2341	0.2377	0.2520	
	0.30	0.1404	0.1385	0.1367	0.1806	0.1736	0.1916	
	0.40	0.0984	0.1007	0.0962	0.1441	0.1440	0.1568	
	0.50	0.0668	0.0655	0.0643	0.0974	0.0966	0.1108	
	0.60	0.0425	0.0401	0.0403	0.0603	0.0607	0.0642	
	0.70	0.0235	0.0219	0.0225	0.0387	0.0377	0.0393	
	0.80	0.0105	0.0104	0.0103	0.0140	0.0135	0.0153	
	0.90	0.0092	0.0091	0.0091	0.0082	0.0059	0.0092	
	1.00	0.0092	0.0091	0.0091	0.0082	0.0059	0.0092	
Non-interpolated Average Precision		0.1036	0.1038	0.1027	0.1266	0.1265	0.1399	

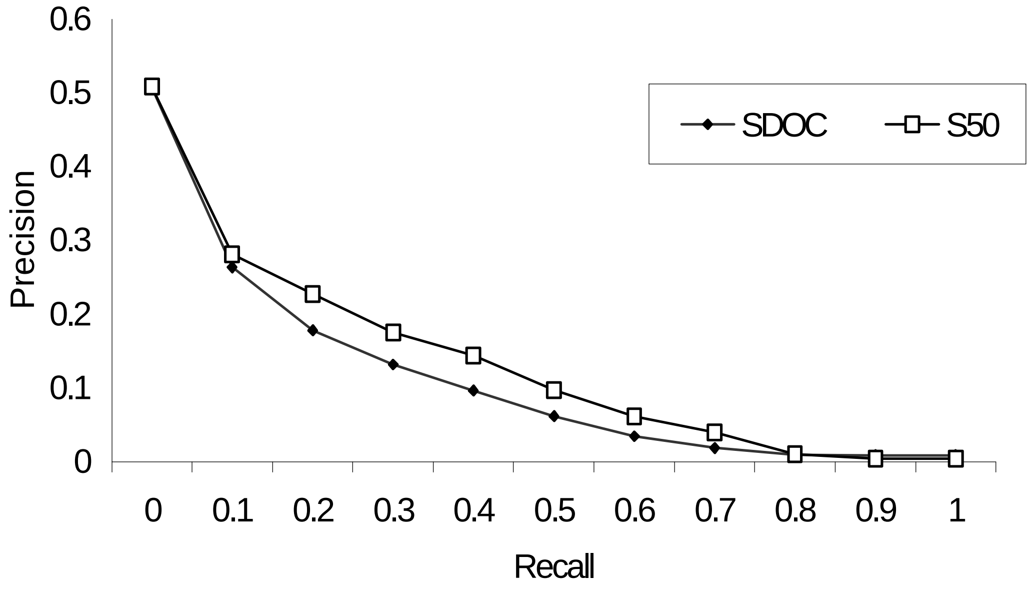

Figure 2. Recall-precision graph for short queries

4.2 Experiment 2: using the highest window score for each document

The best passage to extract from each document may vary in size from document to document. In this experiment, we investigated whether identifying the window size with the highest score for each document and adopting this as the document score improves retrieval effectiveness. As explained earlier, before identifying the window size with the highest score, the window scores were first normalized by taking the $\log_2$ of the scores and then converting them to a z-score with respect to that window size.

Initially, we did not find that taking the highest window score for each document improved retrieval results. We then noticed that the window size of 500 and 800 were most often selected. So, we restricted the window sizes selected to the range 50 to 400. The retrieval results for this are given in Tables 1 and 2 in the column “highest window score.” This approach yielded a retrieval improvement of 4.8% over the size-50 window for long queries, and 5.6% for short queries. The improvements were found to be significant at the 0.05 level for long queries, but not for short queries.

We combined the highest window score with the whole-document score but found that this did not improve the retrieval results.

4.3 Experiment 3: using the best window size for each query

On examining the retrieval results for individual queries, we found that for different queries, different window sizes gave the best retrieval result. The best retrieval results that could be obtained by selecting the best window size for each query are given in Tables 1 and 2 in the column “best window size for query.” The results indicate that a maximum retrieval improvement of 42% can be achieved over whole-document retrieval, if we could predict accurately the best window size to use for each query

The problem is how to identify the best window size for each query. We have not yet found a way to identify the best window size without relevance feedback. However, we attempted the following procedure for selecting a window size for a query using relevance feedback:

- Identify the top 10 documents retrieved by using each window size
- Calculate the retrieval precision for each window size based on the top 10 documents
- If 1 or more window sizes obtain precision values above 0:
    - Then: select the window size that gives the highest retrieval precision
    - Else:
        - identify the rank of the first relevant document retrieved by each window size
        - select the window size for which the first relevant document has the smallest rank.

The retrieval results for using this procedure to select a window size for each query are given in Tables 1 and 2 in the column “predicted best window size for query.” The non-interpolated average precision obtained was about the same as that obtained using the highest window score for each document.

Conclusion

This study explored how window-based passage retrieval scores can be used to improve document retrieval effectiveness in the context of a vector-based retrieval system. The passages were extracted by sliding a window of a certain size across the document, and thus each passage comprises a fixed number of contiguous words from the document. For a particular window size, the highest passage score for the document is taken as the document score for that window size. We examined three ways of using window passage scores:

- using a fixed window size for all queries
- using the highest window score for each document
- using the best window size for each query.

We found that using a fixed window size of 50 gave the best results for the TREC 5 and 6 test collection. This yielded a significant improvement of 24% compared to using the whole-document retrieval score. However, combining the retrieval score for this window size and the whole-document retrieval score did not yield a retrieval improvement.

Identifying the highest window score for each document (for windows varying from 50 to 400 words), and adopting it as the document retrieval score yielded a retrieval improvement of about 5% compared with using the fixed window size of 50 words.

It was observed that different window sizes worked best for different queries. If we could select the best window size for each query, the maximum retrieval improvement that could be obtained over the baseline (taking the whole-document retrieval score) was 42%. However, we have not found an effective way of predicting which window size would give the best results for each query.

Callan’s (1994) experimental results suggest that different window sizes worked best for different corpora. We are currently carrying out experiments to identify the best window sizes to use for the different corpora in the TREC document collection. We are also using logistic regression analysis to investigate how the different window scores can be combined to generate a composite document retrieval score, and how these window scores interact with different factors in determining document relevance.

References

Callan, J.P. (1994). Passage-level evidence in document retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 302-309. New York: ACM.

Frakes, W., & Baeza-Yates, R. (Eds.). (1992). *Information retrieval: Data structures and algorithms*. Englewood Cliffs, NJ: Prentice-Hall.

Kaszkiel, M., & Zobel, J. (1997). Passage retrieval revisited. In *Proceedings of the twentieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 178-185. New York: ACM.

Harman, D. (1992). Ranking algorithms. In W. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 363-392). Englewood Cliffs, NJ: Prentice-Hall.

Hearst, M.A., and Plaunt, C. (1993). Subtopic structuring for full-length document access. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 59-68. New York: ACM.

Mittendorf, E., and Schauble, P. (1994). Document and passage retrieval based on hidden Markov models. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 318-327. New York: ACM.

Moffat, A., Sack-Davis, R., Wilkinson, R., and Zobel, J. (1994). Retrieval of Partial Document. In *Proceedings of the Second Text Retrieval Conference (TREC-2)*, 181-190. NIST Special Publication 500-215.

Salton, G., Allan, J., & Buckley, C. (1993). 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*, 49-58. New York: ACM.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513-523.

Wilkinson, R. (1994). Effective retrieval of structured documents. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 311-317. New York: ACM.

Zobel, J., Moffat, A., Wilkinson, R., and Sack-Davis, R. (1995). Efficient retrieval of partial documents. *Information Processing & Management*, 31(3).