Effective Ranking with Arbitrary Passages

Marcin Kaszkiel
Department of Computer Science, RMIT University, GPO Box 2476V, Melbourne 3001, Australia

Justin Zobel
Department of Computer Science, RMIT University, GPO Box 2476V, Melbourne 3001, Australia

Text retrieval systems store a great variety of documents, from abstracts, newspaper articles, and Web pages to journal articles, books, court transcripts, and legislation. Collections of diverse types of documents expose shortcomings in current approaches to ranking. Use of short fragments of documents, called passages, instead of whole documents can overcome these shortcomings: passage ranking provides convenient units of text to return to the user, can avoid the difficulties of comparing documents of different length, and enables identification of short blocks of relevant material among otherwise irrelevant text. In this article, we compare several kinds of passage in an extensive series of experiments. We introduce a new type of passage, overlapping fragments of either fixed or variable length. We show that ranking with these arbitrary passages gives substantial improvements in retrieval effectiveness over traditional document ranking schemes, particularly for queries on collections of long documents. Ranking with arbitrary passages shows consistent improvements compared to ranking with whole documents, and to ranking with previous passage types that depend on document structure or topic shifts in documents.

Introduction

Documents available in digital form are generated in vast quantities every day, and new methods are required to manage, store, and access them. In particular, locating those that best match a particular interest can be difficult. A suitable access method for full-text databases is to express the information need as a free-text query, which is a description of the information need in natural language or as a list of words. The matching process for free-text queries is to use a heuristic function (or *similarity measure*) that estimates how relevant each document is to the query, based on the shared words in the query and document, and on assigned weights for each word.

Received December 14, 1999; revised August 28, 2000; accepted August 28, 2000.

An alternative access method, which is the topic of this article, is to regard each document as a set of *passages*, where a passage is a contiguous block of text. Instead of computing the similarity of each document to a query, a similarity is computed for each passage (Callan, 1994; Hearst & Plaunt, 1993; Mittendorf & Schäuble, 1994; Salton, Allan, & Buckley, 1993; Wilkinson, 1994; Zobel & Moffat, 1995). The units retrieved can then be the documents from which the most similar passages are drawn—so that passages provide an alternative mechanism for document ranking—or can be the passages themselves. Passage-level access has several advantages over document-level access. First, if passages are relatively short, they embody locality: if query words occur together in the passage, they must be fairly close to each other. Second, passages are more convenient units for viewing and transmission than long documents, and, moreover, in cases such as databases of transcripts, there may be no clear separation of the text into discrete parts; that is, the concept of "document" may not even apply. Third, when used as a mechanism for document retrieval, passages can avoid the difficulties of discrimination between documents of varying lengths. Some similarity measures tend to favor short documents, and thus can be ineffective for collections of documents of mixed lengths (Singhal, Buckley, & Mitra, 1996; Singhal, Salton, Mitra, & Buckley, 1996), whereas, for passages of uniform length, the problems of discrimination between documents of different lengths are less significant. Fourth, for presentation to a user, a short relevant piece of text may be more appropriate than a complete long document.

Many types of passages have been proposed. Some passage types rely on the structural properties of documents such as *sentences*, *paragraphs*, and *sections* (Hearst & Plaunt, 1993; Salton et al., 1993; Wilkinson, 1994; Zobel et al., 1995). Each of these individual structures are considered as passages or are used as building blocks for larger passages. Other passage types are based on *topics* derived by segmenting documents into single-topic units (Beeferman, Berger, & Lafferty, 1997; Hearst, 1994; Ponte & Croft,

1994; Reynar, 1994; Richmond, Smith, & Amitay, 1997; Salton, Allan, & Singhal, 1996; Salton, Singhal, Mitra, & Buckley, 1997). Yet other passage types are based on fixed-length blocks (Callan, 1994; Stanfill & Waltz, 1992). The individual results reported in the literature show that passage-level access is of benefit in full-text databases. One of the outcomes of this article is an evaluation of the effectiveness of different passage types in a common test environment.

Our experimental results compare the effectiveness of several passage types, which are evaluated in terms of their ability to identify relevant documents, that is, documents are retrieved based on the relevance of their passages. We find that use of these types of passages can improve retrieval effectiveness compared with document ranking, by around 50% for some passage types. The effectiveness improvements achieved by the use of passages are significant for databases for which the variability of document length is large, but for databases with uniform document length the improvement is smaller. Nonetheless, for all databases tested, retrieval effectiveness with passage ranking is not usually inferior when compared to document ranking. However, our tests across five different text databases and two different sets of queries show inconsistent retrieval performance for different types of passage. For example, for short queries and text databases of long documents, passages using structural properties of documents are best, whereas, for text databases of uniform document length, only passages that ignore structural properties result in improved retrieval effectiveness.

We propose *arbitrary passages*, which are independent of any structural or semantic properties. Extending our previous work on arbitrary passages (Kaszkiel & Zobel, 1997), we show that document retrieval using fixed-length arbitrary passages is more effective in all cases than whole-document ranking, and that retrieval effectiveness is consistent for a reasonable range of passage lengths. For text databases of uniform document length, where previous passage types had little impact on retrieval effectiveness, fixed-length passages can show significant improvement. Furthermore, comparing the results for individual queries shows that retrieval using fixed-length passages reduces the number of queries with decreased retrieval effectiveness, in contrast to other definitions of passage.

Analysis of our experiments with fixed-length arbitrary passages shows that use of a single passage length can lead to inconsistent performance. As a consequence, we propose an extension, *variable-length arbitrary passages*, by relaxing the restriction on passage length. As the query is processed, several passage lengths are considered. When the processing of each document is complete, the best passage of any length is selected. We show that variable-length arbitrary passage ranking improves effectiveness compared with fixed-length arbitrary passage ranking, by 2 to 9%. This improvement is at the expense of additional computation required to process a large number of passage lengths; however, in other work we have developed efficient algorithms for ranking arbitrary passages, showing that it is practical on realistic collections (Kaszkiel, Zobel, & Sacks-Davis, 1999). We use significance tests to examine the validity of all results, and show that, for much of our test data, ranking with variable-length arbitrary passages is clearly superior to whole-document ranking.

Background

Similarity Measures

There are several different models that provide a basis for matching full-text documents to free-text queries, in particular the *vector-space* (Salton & Buckley, 1988; Salton & McGill, 1983) and *probabilistic* (Crestani, Lalmas, van Rijsbergen, & Campbell, 1998; Robertson & Walker, 1994; van Rijsbergen, 1979) models. Many similarity measures have been proposed and investigated, but no single function is significantly superior to others (Salton & Buckley, 1988; Zobel & Moffat, 1998); relative performance can vary significantly, depending on the database and the set of queries. An effective vector-space similarity measure is the *cosine measure*, for which one formulation for computing the similarity of a document d to query q is (Zobel & Moffat, 1998):

$$C(q, d) = \frac{\sum_{t \in q \wedge d} (w_{q,t} \cdot w_{d,t})}{W_d \cdot W_q} \quad (1)$$

with:

$$W_{d,t} = \log_e(f_{d,t} + 1),$$

$$w_{q,t} = \log_e(f_{q,t} + 1) \cdot \log_e\left(\frac{N}{f_t} + 1\right),$$

$$W_d = \sqrt{\sum_{t \in d} w_{d,t}^2},$$

$$W_q = \sqrt{\sum_{t \in q} w_{q,t}^2},$$

where $f_{x,t}$ is the number of occurrences or *frequency* of term t in x; there are N documents; f_t is the number of distinct documents containing t; and the expression $\log_e(N/f_t + 1)$ is the *inverse document frequency*, a representation of the rareness of t in the collection. The quantity $w_{x,t}$ is the weight of term t in query or document x and W_x is a representation of the length of x.

A variant form is the pivoted-cosine measure (Singhal et al., 1996a; Singhal, Choi, Hindle, Lewis, & Pereira, 1998), which is designed to remedy the problems associated with the document length normalization component W_d in Equation 1; one shortcoming of the cosine measure is that it

TABLE 1. Statistics for five text collections used in the experiments.

	FR-12	FR-24	TREC-24	TREC-45	WSJ-12
Number of documents	45,820	75,490	524,929	556,077	173,252
Text size (Mb)	469	604	2,059	2,134	488
Dictionary entries	140,227	166,824	697,593	716,594	156,796
Longest document (Kb)	2,577	6,245	6,245	6,245	133
Median doc length (Kb)	3.4	5.8	2.5	2.5	1.8
Average doc length (Kb)	10.5	8.2	4.0	3.9	3.0
Short queries (avg)	3.4	3.2	3.2	2.5	3.3
Long queries (avg)	39.3	27.5	30.4	26.2	44.1

favors short documents over long. With this measure, the similarity between document d and query q can be computed as:

$$\text{sim}(q, d) = \sum_{t \in q \wedge d} \left(\frac{w_{q,t} \cdot w_{d,t}}{W_d} \right) \quad (2)$$

where q is a query, d is a document,

$$w_{q,t} = 1 + \log_e(1 + \log_e(f_{q,t})) \cdot \log_e\left(\frac{N+1}{f_t}\right),$$

$$w_{d,t} = 1 + \log_e(1 + \log_e(f_{d,t})),$$

$$W_d = (1 - \text{slope}) + \text{slope} \cdot \frac{d_{len}}{avr - d_{len}}.$$

The value d_{len} is a document length in raw bytes and avr_d_{len} is the average document length in the collection. Slope changes the cosine normalization factor; the value of 0.2 is used throughout this article (Singhal, 1997; Singhal et al., 1998). The overall effect is to skew the normalization in favor of long documents, with the degree of skew controlled by *slope*.

The pivoted-cosine measure has consistently been shown to be superior at the Text REtrieval Conference (TREC) (Vorhees & Harman, 1997, 1998). A probabilistic approach that is of similar effectiveness is the Okapi measure developed by the City University group (Robertson & Walker, 1994; Robertson, Walker, & Beaulieu, 1998; Walker, Robertson, Boughanem, Jones, & Sparck-Jones, 1997). However, all our experiments are in the vector-space model. In experiments with these measures, we have found that the difference between them is statistically insignificant.

Test Data

Test collections are used to evaluate and compare different retrieval systems (Salton & McGill, 1983). We use the large test collections built as part of the TREC initiative (Harman, 1995). TREC includes heterogeneous data, and the lengths of documents vary from tens of bytes to a few megabytes. In TREC, queries are represented in the form of *topics* that describe the information need at different levels. Each topic consists of three fields: "title," "description," and "narrative." In our experiments, we use two types of queries: *short* and *long*. The short queries include words from title fields, and the long queries are the full topics. For the Internet most queries are short, typically around four words or less (Lu & Keefer, 1994). Longer queries are used by experienced users to describe information needs in greater detail. The intention of using both types of queries is to demonstrate the different characteristics of ranking when used with short and long queries.

We use five test collections. The first two text collections, FR-12 and FR-24, correspond to an environment of long documents, with a large variance in the document length. These collections are the Federal Register data from disks 1 & 2 and disks 2 & 4, respectively. For such text collections, with a large spectrum of document lengths, whole-document ranking is expected to perform poorly. The next two collections, TREC-24 and TREC-45, are more heterogeneous. They are the full contents of TREC disks 2 & 4 and disks 4 & 5, respectively. The last text collection, WSJ-12, is of similar magnitude to the first two collections but contains substantially shorter documents; it is the Wall Street Journal data from disks 2 & 4. Details of these collections are shown in Table 1.

Retrieval Effectiveness

A common benchmark used to measure retrieval effectiveness is *precision* and *recall* (Salton, 1989). Typical standard recall levels, referred to as 11-point levels, are 0, 10, . . . , 90, and 100%. Results can be summarised as a single value, the average precision across the 11-point recall levels. In this article, we use the average 11-point precision, and precision at 5, 10, 20, 30, and 200 document cutoffs, to compare retrieval systems.

Measuring differences in precision and recall between retrieval systems is only indicative of the relative performance. It is also necessary to establish whether the difference is statistically significant. Per-query recall-precision figures can be used in conjunction with statistical significance tests to establish the likelihood that a difference is significant. We use a nonparametric test, the *Wilcoxon signed rank test*, which has been shown by Zobel (1998)

(and others) to be suitable for this task. In our comparisons, a 95% level of confidence is used to find whether the results are statistically significant.

Passages in Information Retrieval

Documents can be accessed by using their content for matching and retrieval. Entire documents are treated independently, each represented by the terms selected during automatic indexing. Queries are matched against the document representation. However, this approach has disadvantages. For example, when a long document is retrieved, it is difficult to present it to the user, and it may not be desirable to retrieve a document that is long and not entirely relevant. Ideally, users should be guided to the relevant section of the document.

Another problem is that a long, relevant document may be lowly ranked, for several reasons. First, in contrast to concise documents such as abstracts—in which most words are specific terms that accurately describe the main topics and discriminate well between relevant and irrelevant documents (Salton & McGill, 1983)—a long document may consist of many thousands of words.

Second, most text database systems treat documents as bags of words, ignoring relative word positions in documents. This has an implication for document ranking. For example, consider a query "space travel." Documents that discuss "seating space" and "international travel" could be retrieved but are not relevant.

Third, many widely used similarity measures have been shown to favour short documents (Singhal et al., 1996a). However, long documents that have only a small relevant fragment have less chance of being highly ranked than shorter documents containing a similar text fragment, although Singhal et al. (1996a) showed that, in the TREC data, long documents have a higher probability of being relevant than do short documents. Analysis of text databases has shown that similarity functions can be adjusted to help remedy the problems of document length differences (Singhal et al., 1996b; Walker et al., 1997).

Another solution that deals with document length normalization is to summarize and then use the summaries for ranking. In automatic text summarization, or abstracting, the key text components are extracted to represent each document (Brandow, Mitze, & Rau, 1995; Paice & Jones, 1993; Salton et al. 1997). However, ranking with text summarization may not identify a document in which only a fragment is relevant, and the text used for measuring similarity may have been scattered through the original document.

Passages

An alternative approach to matching whole documents is to consider each document as a set of passages. A *passage* is any sequence of text from a document. Query evaluation proceeds by identifying the passages in the document collection that are most similar to the query. Then either the documents containing the highest ranked passages are returned to the user, or the passages are returned together with context information such as the titles of the documents and information about the location of the passages within the documents' structures.

Passage retrieval has several potential advantages in contrast to whole document retrieval. First, because passages are relatively short, they embody locality: if the query terms occur together in the passage they must be fairly close to each other. Second, passages are more convenient to the user than long documents. In some instances, such as collections of transcripts, there may be no clear separation of the text into discrete parts and, therefore, the concept of document does not apply. As another example, in a database of the full text of books, it is not clear whether a book or a chapter would be considered as a document. Third, when passages are used as a mechanism for document retrieval, they can avoid the difficulties of document length normalization; for passages of equal length the problems of normalization are not significant. Finally, it can be argued that a document that has a short passage containing of a high density of words that match a query is more likely to be relevant than a document with no such passage, even if the latter contains a reasonable number of matching words across its length and has higher overall similarity.

Experimental evidence suggests that document ranking based on passages may be more effective than ranking of entire documents. Hearst and Plaunt (1993) showed that extracting the best passages from a document and adding scores for several passages produces better ranking than that based on whole-document scores. Callan (1994) showed that ordering documents based on the score of the best passage may be up to 20% more effective than a standard document ranking. Salton et al. (1993) used passages to filter out documents with low passage scores, showing that, by restricting the retrieval to those documents that have high document and high passage similarity, retrieval improved by up to 22.5% compared with standard ranking.

Representing a document by a single passage is not the only option. A longer document with several highly significant passages would be disadvantaged because only a single passage is used to represent it. An extension to single passage ranking is to consider several passages for each document and use them to represent the document. Hearst and Plaunt (1993) used the sum of several passages to relate the similarity of documents to queries, which was more effective than single passage ranking. Clarke, Cormack, and Burkowski (1995) developed a *shortest substring segments* approach using Boolean queries to match document segments. Only those segments that satisfy the Boolean expression are considered. Individual segments are ranked by the inverse of the absolute length and documents are ranked by the sum of the scores of nonoverlapped segments matched in the document. Yet another approach is to combine passage similarities with document similarities (Buckley, Salton, Allan, & Singhal, 1994; Callan, 1994; Wilkinson,

1994). For example, Callan (1994) combined two raw similarities: an entire document and a best passage (a window). Buckley et al. (1994) used a slightly more complex function to combine best passage score with document score.

Passage retrieval also has other applications. Cormack et al. (1997, 1998) used short document segments, around 20 words in length, as passages. Retrieved passages were used to assess documents as being either relevant or nonrelevant. The samples of documents judged were as accurate as the official judgments (Voorhees & Harman, 1998; Walker et al., 1997), strongly suggesting that short passages can be used to indicate relevance. Hearst (1994) and Plaunt's TextTiling algorithm (1993) partitions full-length documents into multiparagraph units to approximate a document's subtopic structure. Such an approach is particularly useful when document structure is absent or does not reflect the text content. Passages can also be used in relevance feedback and automatic query expansion. The effectiveness of automatic query expansion is degraded when long documents are used (Allan, 1995); instead, only the part of the document that is most similar to the query should be used for feedback. Allan (1995) and Xu and Croft (1996) showed that using passages instead of full-text documents in automatic query expansion can improve the retrieval effectiveness of queries, and passages have also been used in other work with relevance feedback (Cormack et al., 1997; Papka & Allan, 1997).

Many types of passages have been successfully used to retrieve documents. These types can be classified as either discourse, semantic, and windows (Callan, 1994). In previous research, experimental results demonstrating the effectiveness of passage retrieval have been obtained in diverse test environments. However, there has been no comprehensive study that compares a large number of types of passages, and the results reported in the literature are not directly comparable because different test collections were used. One of our objectives was to study existing passage types in a uniform test environment.

Discourse passages

Documents usually have structural or logical divisions such as sentences, paragraphs, and sections, marked up in standards such as XML. The discourse (or logical) components of documents can be regarded as passages (Callan, 1994; Hearst & Plaunt, 1994; Lalmas & Ruthven, 1997; Salton et al., 1993; Wilkinson, 1994). This definition of passage is intuitive, because sentences should convey a single idea; paragraphs should be about one topic; and sections should be about one issue.

A problem with discourse passages is that they require a high degree of consistency between authors. Callan (1994) observed that the structure of a document might be unrelated to its content, because documents can be structured in a particular way simply for presentation. Also, even though most documents are supplied with their structure, manual processing is required for those without it, thus making discourse passages impractical, as can be the case when a document is the output of a speech recognition system (Ponte & Croft, 1997). Another problem with discourse passages is that their length can vary, from very long to very short. In addition, long passages are likely to include more than one topic; retrieving long passages contradicts one of the main aims of passage retrieval.

Semantic passages

An alternative approach is to segment documents into *semantic* passages, corresponding to the topical structure of documents (Beeferman et al., 1997; Hearst, 1994; Ponte & Croft, 1997; Reynar, 1994; Richmond et al., 1997; Salton et al., 1996, 1997). The principal idea is to partition documents into segments, each corresponding to a topic or to a subtopic. It is, therefore, attractive to develop algorithms that derive segments based on topic or semantic properties. Several such algorithms have been developed. Reynar (1994) proposed an algorithm that locates semantic boundaries based on detection of repetition of lexical items such as words or phrases. Beeferman et al. (1997) used short- and long-term statistical models that keep track of word occurrence patterns, near and far from the current position in text, to locate topic changes, and also used lexical hints such as sentence and paragraph boundaries. Yaari (1997) applied a hierarchical agglomerative clustering algorithm to partition full-text documents, which is similar to a technique used by Maarek and Wecker (1994). The algorithm incrementally joins adjacent paragraphs on the basis of their similarity. Salton et al. (1996, 1997) derived text segments that helped with summarizing documents by computing similarities between text paragraphs. Ponte and Croft (1997) developed an algorithm that segments texts into short topics, assumed to be about three sentences long.

An algorithm that is well-suited to passage retrieval from large collections such as the TREC data is that of Hearst (1994), known as TextTiling, which partitions full-text documents into coherent multiparagraph units. This scheme creates a subtopic structure for a document using multiparagraph segmentation. Single-paragraph passages are avoided because topics can be discussed in consecutive paragraphs. The algorithm relies on word frequencies to recognise topic shifts. Richmond et al. [28] extended the TextTile algorithm by introducing a new measure of word significance, which uses the relative occurrence of words in documents to compute the scores between adjacent blocks. Experimental results suggest that the extended algorithm is slightly more reliable than the original TextTile algorithm.

In this article, Hearst's algorithm is used to determine semantic passages. The first step of the algorithm is to tokenise the input by recognizing words, removing words with low content, and creating token-sequences, which are nonoverlapped sequences of words. Token-sequences are used in place of sentences. Token-sequences are too short to be reliably compared with each other. Instead, blocks are created from k consecutive token-sequences. The blocks

highly overlap with each other. The similarities between adjacent blocks are computed to form a gap score between adjacent blocks. Adjacent gap scores are used to derive locations at which topic shifts are most probable. Each topic shift determines the end of a semantic passage. Each semantic passage is referred to as a *tile.*

Regardless of the segmentation technique, an advantage of semantic passages is that they can be applied even where the logical structure of documents is not explicit. This can be useful when, for example, documents have been created using OCR or speech recognition technology. Discovering semantic passages is computationally expensive, but this cost is only incurred once. However, the accuracy of segmentation compared with human segmentation is not yet perfect (Hearst, 1994; Richmond et al., 1997).

Window-Based Passages

Structural properties of documents are not always explicit, retrieval requirements vary depending on the user need, and semantic segmentation can be inaccurate. An alternative to discourse and semantic passages is to break documents into passages of fixed length, often referred to as nonoverlapped windows. If paragraph boundaries are known, they can be considered, but if they are not available, then passages can simply be defined as sequences of words.

The passage should be in a fixed range of sizes based on number of words, not too long or too short. Callan (1994) used a word-based approach, by defining a passage, or a *window,* as a fixed-length sequence of words. Zobel et al. (1995) and Callan (1994) considered paragraphs instead of words as the basic unit, and used heuristics to bound their lengths. For instance, short paragraphs are merged with subsequent paragraphs, and paragraphs longer than some minimum length are kept intact. Zobel et al. (1995) referred to such passages as *pages,* because they approximate a physical page of text. Each page is around 2 kilobytes. Stanfill and Waltz (1992) define a passage as a block of 30 words, and segment documents into sequential blocks. Consecutive blocks can be joined into a larger text segment, to address the problems of retrieving blocks of texts that are too short.

The main advantage of window-based passages is that they are easy to construct, irrespective of the text. However, there are disadvantages. If window-based passages are retrieved and presented to the user, they are likely to be confusing unless additional information is presented, describing the context from which the passage has been selected; and window-based passages are static, because, once they are defined, they are also indexed. However, Callan (1994) and Kaszkiel and Zobel (1997) suggested a more dynamic definition of windows, discussed later.

Experimental Results

We experimentally evaluated some of the passage types given above, according to their ability to identify relevant documents. The similarity of the best passage from each document was used to represent the score of the document, and documents were ranked according to their similarity scores. However, document retrieval based on single passages might not be the best technique for evaluating passages. Other techniques, such as using several passages per document or combining the best passage score with the document score, could also be used (Callan, 1995; Clarke et al., 1995; Hearst & Plaunt, 1993). However, these techniques introduce additional variables, such as how many passages to use per document or how to combine them. We believe that document retrieval based on a single best passage is a sound evaluation metric for discriminating between different passage types because it reduces the number of parameters involved. Another way of evaluating passage retrieval would be to retrieve each passage and manually assess the relevance of the passage, instead of the whole document, but pragmatically this is difficult. In all the experiments below, documents are retrieved using either whole-document ranking or passage ranking.

In the experiments, at least one passage type is used from each of the three categories discussed above. Discourse passages used in these experiments are *paragraphs* and *sections,* referred to as PARAGRAPHS and SECTIONS. They directly correspond to paragraphs and sections as marked explicitly in documents or determined from common conventions, such as blank lines between paragraphs. We consider sentences too short for estimating the relevance of documents. Evaluation of semantic passages is based on the TextTile algorithm (Hearst & Plaunt, 1993), which was used as a baseline approach in other topic segmentation algorithms. These passages are referred to as TILES. The source code for the TextTile algorithm was made available by Hearst.[1]

There are two types of window-based passages: those that do not consider logical structure of documents (Callan, 1994), referred to as WINDOWS, and those that restrict windows to minimum length based on some limited document structure such as paragraph boundaries (Zobel et al., 1995), referred to as PAGES. Past results, including our own experiments, showed that an effective length for WINDOWS is any size between 150 and 350 words (Callan, 1994; Kazkiel & Zobel, 1997), hence the use of WINDOWS-150 and WINDOWS-350. Documents are partitioned into nonoverlapped windows, with the first WINDOWS-150 starting at the first word of a document, the second WINDOWS-150 starting at the 151st word of a document and so on. In contrast to WINDOWS, PAGES considers paragraph information and bounds them by a minimum length (Zobel et al., 1995). Merging short consecutive paragraphs avoids having paragraphs that are one or two sentences long. Experiments on the TREC data (Zobel et al., 1995) showed that PAGES of about 2,000 characters (or bytes) are best.

[1] The TextTile implementation can be downloaded from: http://elib.cs.berkeley.edu/src/texttiles.

TABLE 2. "Cosine" experiments with the FR-12 collection, 26 queries selected from 51–100.

	Precision at *N* documents						
	5	10	20	30	200	AvgP	%Δ
Short queries							
Document	0.0857	0.0857	0.0667	0.0603	0.0267	0.1283	0.0
PARAGRAPHS	0.1333	0.1143	0.0881	0.0762	0.0310	0.2327	+81.4
SECTIONS	0.1238	0.0762	0.0500	0.0508	0.0229	0.1245	−3.0
TILES	0.1238	0.1238	0.1048	0.0968	0.0367	0.1985	+54.7
PAGES	0.1810	0.1429	0.1143	0.0905	0.0355	0.2250	+75.4
WINDOWS-150	0.1524	0.1333	0.1095	0.0952	0.0383	0.2580	+101.1
WINDOWS-350	0.1714	0.1476	0.1214	0.1063	0.0383	0.2701	+110.5
Long queries							
Document	0.2286	0.2238	0.1762	0.1508	0.0712	0.2928	0.0
PARAGRAPHS	0.2857	0.2190	0.1905	0.1683	0.0645	0.2790	−4.7
SECTIONS	0.1238	0.1286	0.1310	0.1095	0.0452	0.0948	−67.6
TILES	0.2667	0.2286	0.1952	0.1714	0.0674	0.2358	−19.5
PAGES	0.2857	0.2524	0.2119	0.1810	0.0681	0.3143	+7.3
WINDOWS-150	0.2857	0.2381	0.1952	0.1714	0.0679	0.3127	+6.8
WINDOWS-350	0.2952	0.2571	0.2214	0.1841	0.0705	0.3245	+10.8

All of these passage types are defined at indexing time, are indexed as independent units, and are nonoverlapped, with no text shared between any of the passages. All tables that report experimental results include average precision values for five document cutoffs: 5, 10, 20, 30, and 200. In addition, the overall effectiveness is summarized as an average precision, AvgP. The differences between systems, %Δ, are based on the average precision.

FR-12 collection

The first experiment is with the Federal Register documents from disks 1 and 2 of the TREC data (FR-12). The query set consisted of 26 topics from 251–300 that have at least one relevant document in FR-12. It is on this kind of collection that passage retrieval should yield the greatest improvements compared with whole-document retrieval. With the cosine measure shown to be biased towards short documents, passage ranking should diminish the problem because passage lengths are less diverse than document lengths.

Using a single passage to retrieve documents can be much superior to using whole-document ranking with the cosine measure, as shown in Table 2. WINDOWS demonstrate the largest improvements, probably because their length range is not as skewed as that of PARAGRAPHS, SECTIONS, or TILES. PARAGRAPHS do not perform well because many of them are very short. The results for long queries are different. When the query describes the information need in a narrative manner, the mismatch in the length between document and a query is not as significant. As the query gets longer, short passages, such as some of the semantic or discourse passages, do not capture enough information to differentiate well between relevant and nonrelevant documents. However, using window-based passages, some improvements occur. Most passage types have higher precision at low document cutoffs than do whole documents, so that on average a user sees more relevant documents in the top ranks.

On collections such as FR-12, where documents are of mixed lengths (from 93 bytes to 2.5 Mb), document ranking using the cosine measure is expected to perform poorly (Callan, 1994). To improve retrieval effectiveness, pivoting can be used (Singhal et al., 1996a) (see Equation 2). We repeated the same set of experiments using the pivoted cosine measure, except for WINDOWS, whose lengths are more or less equal. Table 3 shows these results. The improvements in the average precision are shown with respect to whole-document ranking using the pivoted-cosine measure.

The results of whole-document ranking are mixed. For short queries, the pivoted-cosine measure improves over the cosine measure by 61% in average precision. However, for long queries, the pivoted-cosine measure degraded the retrieval effectiveness by 17% compared with the cosine measure. This implies that the cosine measure produces better retrieval results than pivoted-cosine measure when queries are long and document lengths vary. The long versions of the 21 queries in this test collection average 39 words each, however, and such detailed and narrative descriptions of information needs may not be common in practice.

Using the pivoted-cosine measure to rank predefined passages such as PARAGRAPHS, SECTIONS, TILES, or PAGES, shows consistent improvements over ranking based on the cosine measure. The largest improvements are for SECTIONS, where the average precision is doubled, an increase that is unsurprising because the lengths of SECTIONS are highly variable. For other predefined passages the improvements are still substantial, averaging 20% per passage type. The results for passage-based ranking with variable lengths sup-

TABLE 3. "Pivoted-cosine" experiments with the FR-12 collection, 26 queries selected from 51–100.

	Precision at *N* documents						
	5	10	20	30	200	AvgP	%Δ
Short queries							
Document	0.1619	0.1429	0.1024	0.0937	0.0329	0.2075	0.0
PARAGRAPHS	0.1524	0.1333	0.1095	0.1079	0.0405	0.3001	+44.6
SECTIONS	0.1524	0.1333	0.1024	0.0873	0.0340	0.2638	+27.1
TILES	0.1714	0.1238	0.1119	0.1000	0.0381	0.2502	+20.6
PAGES	0.2095	0.1667	0.1238	0.1079	0.0400	0.3067	+47.8
WINDOWS-150	0.1524	0.1333	0.1095	0.0952	0.0383	0.2580	+24.3
WINDOWS-350	0.1714	0.1476	0.1214	0.1063	0.0383	0.2701	+30.2
Long queries							
Document	0.1810	0.1571	0.1357	0.1365	0.0533	0.2417	0.0
PARAGRAPHS	0.3048	0.2571	0.2143	0.1905	0.0700	0.3616	+49.6
SECTIONS	0.2667	0.2286	0.1857	0.1571	0.0631	0.2010	−16.8
TILES	0.3238	0.2571	0.2143	0.1810	0.0681	0.2674	+10.6
PAGES	0.3143	0.2524	0.2071	0.1841	0.0731	0.3502	+44.9
WINDOWS-150	0.2857	0.2381	0.1952	0.1714	0.0679	0.3127	+29.4
WINDOWS-350	0.2952	0.2571	0.2214	0.1841	0.0705	0.3245	+34.2

port similar results found when this measure was used for document ranking (Singhal et al., 1996a).

The improvement of predefined passage ranking over whole-document ranking with the pivoted-cosine measure is not as significant as for the cosine measure. For short queries, an average and consistent improvement of over 20% is achieved with passage ranking over document ranking, compared with over 50% improvements in experiments with the cosine measure (see Table 2). The difference between passage and whole-document ranking diminishes due to the better term weighting scheme and length normalization.

In this experiment, documents are retrieved according to entire document content or just the best passage. For collections such as FR-12, where documents can be long, passage retrieval is more appropriate than document retrieval. For documents that were retrieved by document ranking (cosine or pivoted-cosine measure), lengths in the top 30 have typically varied from 1.4 Kb to 0.6 Mb. Average length for the cosine measure was 22 Kb and for the pivoted-cosine measure was 74 Kb. These documents are approximately 10 pages of text. Thus, it is undesirable to return entire documents to the user. When documents are retrieved using the single best passage method, the document length increases to an average of 140 Kb, with the longest being 1.3 Mb.

Based on the results in Table 2 and 3, documents ranked using any type of passage results in more effective retrieval than when whole-documents are ranked using the pivoted-cosine measure.

FR-24 collection

In this experiment the aim was to validate the results achieved on the FR-12 collection, by changing the set of documents and the query set, yet preserving similar document characteristics to FR-12. The documents of the Federal Register from disk 2 and 4 (FR-24) are used together with 26 topics from 251–300 (those topics that have at least one relevant document in FR-24). The pivoted-cosine measure appears to be superior to the cosine measure, and so, from this point on, only pivoted cosine is used for whole-document ranking, and for passage ranking where passage lengths vary. Full results are omitted; summary results are shown later in this section in Table 7. In the FR-24 collection with long queries, however, every passage type other than SECTIONS outperformed whole-document ranking.

TREC-24 collection

The results achieved using document retrieval based on passages with the FR-12 and FR-24 collections are encouraging. However, for large text collections with documents of more uniform size than those in either FR collection, whole-document ranking based on the pivoted-cosine measure is expected to perform reasonably well (Singhal et al., 1996a). We applied the same set of experiments to a larger text collection with more uniform document lengths. Two full disks of TREC data were selected (TREC-24), the test data used for the TREC 5 conference (Voorhees & Harman, 1997). The query set contained 50 topics, numbered from 251 to 300. The pivoted-cosine measure was used for whole-document, PARAGRAPHS, PAGES, and TILES ranking. WINDOWS were ranked with the cosine measure without length normalization. The experimental results for the TREC-24 collection are summarized later in this section.

TREC-45 collection

The next series of the experiments is designed to confirm the results achieved on the previous test collections. The test data was that used for the TREC 6 conference (Voorhees &

TABLE 4. "Pivoted-cosine" experiments with the TREC-45 collection, queries 301–350.

	Precision at *N* documents						
	5	10	20	30	200	AvgP	%Δ
Short queries							
Document	0.4160	0.3520	0.3180	0.2740	0.1184	0.1909	0.0
PARAGRAPHS	0.3920	0.3320	0.2780	0.2413	0.1050	0.1699	−11.0
TILES	0.3960	0.3460	0.2960	0.2560	0.1116	0.1840	−3.6
PAGES	0.4320	0.3520	0.3070	0.2647	0.1159	0.1910	+0.1
WINDOWS-150	0.3320	0.3040	0.2510	0.2293	0.0950	0.1577	−17.4
WINDOWS-350	0.3760	0.3400	0.2850	0.2473	0.1020	0.1719	−10.0
Long queries							
Document	0.5160	0.4240	0.3310	0.2880	0.1234	0.2037	0.0
PARAGRAPHS	0.5000	0.4020	0.3320	0.2840	0.1290	0.2113	+3.7
TILES	0.5040	0.4200	0.3360	0.2953	0.1286	0.2100	+3.1
PAGES	0.4920	0.4320	0.3320	0.2960	0.1324	0.2182	+7.1
WINDOWS-150	0.4240	0.3560	0.3020	0.2673	0.1129	0.1809	−11.2
WINDOWS-350	0.4280	0.3740	0.3190	0.2800	0.1220	0.1859	−8.7

Harman, 1997). Documents from disks 4 and 5 were used (TREC-45). The query set consisted of 50 topics, numbered from 301 to 350. As for the previous collections, the pivoted-cosine measure was used for whole-document, PARAGRAPHS, TILES, and PAGES ranking. WINDOWS were ranked with the cosine measure, and no length normalization employed. Experimental results are shown in Table 4. They show that, for short queries, whole-document ranking is consistently better than the passage retrieval techniques, as was also observed for the TREC-24 collection. Only retrieval based on PAGES achieves a level of effectiveness equivalent to whole-document ranking.

For long queries, ranking based on PARAGRAPHS and TILES is marginally better than that based on document ranking. However, the best results are achieved by PAGES, where the average precision is improved by 7.1%. Also, PAGES has the highest precision at the 10, 30, and 200 document cutoffs. Only PAGES produces a consistent improvement over whole-document ranking.

For a text collection such as TREC-24 or TREC-45, passages are not as attractive as for the FR-12 and FR-24 collections; document retrieval based on whole-document ranking using the pivoted-cosine measure is almost as effective as document retrieval based on passages. The average length of relevant documents in the FR collections, 145 Kb, is 10 times greater than average document length in the collection overall. In the TREC-24 collection, where document lengths are more uniform and most are short, the average length of a relevant document in the TREC-24 collection, 16 Kb, is only three times longer than the average document length. As a consequence, the majority of relevant documents are retrieved with whole-document ranking.

WSJ-12 collection

We used another text collection to test passage ranking in a collection of uniform document lengths. This collection is the Wall Street Journal from disks 1 and 2 (WSJ-12), where almost all documents are shorter than 10 Kb. The query set is topics 51 to 100. The number of relevant documents per query is larger than for the FR-12 collection. Also, because almost all documents are short, the average length of relevant documents is much shorter than those in either the FR-12 or FR-24 collection.

The results, summarized later, demonstrate that in this case the benefit of passages in document retrieval is limited. In three cases out of 10, document retrieval based on passages degrades the effectiveness: WINDOWS-150 and WINDOWS-350 for short queries, and TILES for long queries. In other cases, document ordering based on passages is at least as effective as whole-document ranking. For long queries, there is no benefit in using these kinds of passages.

Significance and Analysis

These results demonstrate that, in terms of typical effectiveness, document retrieval based on passages is up to 50% better than whole-document ranking. In information retrieval experiments, an improvement of over 10% in the average precision is sometimes regarded as significant (Keen, 1992). We evaluated this interpretation of the results with the Wilcoxon signed rank test.

Table 5 shows the percentage of queries where passage-based ranking performed better than whole-document ranking. Independent of the passage type used and the query length, passage-based ranking is better than whole-document ranking for the FR-12 collection. However, comparing passage-based ranking with document ranking for the FR-12 collection, the only statistically significant results are for PAGES and WINDOWS-150 (for short queries) and for PARAGRAPHS and PAGES (for long queries). By considering the average precision improvements in Table 3 (for FR-12 collection), we see that having a large improvement in the average precision, such as for PARAGRAPHS (44.6%), does not necessarily mean that the results are significantly better.

TABLE 5. Comparison of average precision of passage-based versus whole-document ranking.

	PARA	SEC	TILE	PAGE	WIN-150	WIN-350
Short queries						
FR-12	48/33	43/43	48/43	**52/33**	52/33	**52/19**
FR-24	42/42	31/4	35/54	38/46	35/50	35/46
TREC-24	**30/66**	—	**6/92**	44/52	32/62	36/58
TREC-45	**14/82**	—	**28/70**	38/58	**36/62**	38/60
WSJ-12	58/42	—	50/50	**58/40**	**34/64**	58/42
Long queries						
FR-12	**67/24**	57/38	71/24	**76/14**	62/29	71/24
FR-24	35/54	58/38	46/50	38/54	58/38	50/42
TREC-24	32/64	—	52/44	54/42	46/50	52/44
TREC-45	**60/40**	—	64/34	**78/20**	44/56	50/50
WSJ-12	52/48	—	**38/62**	66/34	48/52	54/46

Each entry has two numbers, X and Y (that is, X/Y). X is the percentage of queries where the given passage-based ranking technique is better that whole-document ranking. Y is the percentage of queries where the given passage-based ranking technique is worse than whole-document ranking. The numbers in bold represent the significant results using the Wilcoxon test with a 95% confidence level.

For the FR-24 test collection, which also consists of many long documents, the Wilcoxon test does not reveal any significant differences between whole-document ranking and passage ranking. As Table 5 shows, for most passage types, the retrieval effectiveness for the majority of queries is degraded by passage ranking. It is surprising that, for long queries, the average precision improved by up to 44%, and yet there is no significant difference between passage and document ranking. To add to this contradiction, for PARAGRAPHS, TILES, and PAGES, where the average precision improved by 26, 25, and 36%, respectively, over whole-document ranking, the majority of the queries are less effective!

For the TREC-24 collection, the improvements in the average precision for passage-based ranking over whole-document ranking are mild. For short queries, using the Wilcoxon test, TILES and PARAGRAPHS are significantly worse than whole-document ranking. For other passages, most queries are less effective but no significant difference is detected. For long queries, document retrieval based on TILES, PAGES, or WINDOWS shows consistent improvements over document ranking.

The short queries for the TREC-45 collection show that whole-document ranking is significantly better than passage-based ranking techniques other than PAGES and WINDOWS-350. A surprising result is that retrieval based on TILES is only 3.7% worse than whole-document ranking, and yet the difference between the systems is significant. Similar significance results were reported by Zobel (1998), where thousands of systems are compared using average precision and the Wilcoxon test. For long queries, PAGES and PARAGRAPHS show significant improvement over document ranking, despite only a small increase in the average precision (see Table 4).

For WSJ-12, because most of the documents are short—only a few exceed 4 Kb—a segmentation technique is expected to have only minimal impact on retrieval effectiveness. However, this is not the case. For short queries, passages generally improve the retrieval effectiveness over whole-document ranking, except for WINDOWS-150. Also, PAGES prove to be significantly better than whole-document ranking, even though the improvement in the average precision is just 5.9%. For long queries, passage-based retrieval did not significantly improve on whole-document ranking.

In summary, the Wilcoxon tests on text collections such as FR, where there is a smaller number of queries, show that a large average increase in precision does not necessarily imply significant improvements. For other collections, such as TREC-24 and TREC-45, where there are more queries, the Wilcoxon test is more consistent. Similar results were observed for precision at 20 documents retrieved; passages were usually helpful, sometimes significantly, but, other than with PAGES, were significantly worse for WSJ-12.

Examining the percentage ratios of queries where passages are superior to whole-document ranking, we observe that PAGES showed consistent improvements over whole-document ranking. However, on three occasions PARAGRAPHS slightly outperformed PAGES, while TILES, SECTIONS, and WINDOWS were close to PAGES but did not show significant improvements. To quantify this comparison, we present another table, which uses the PAGES technique as a baseline. Comparisons are summarized in Table 6. An interpretation of this table is as follows: the baseline or PAGES-based ranking is better if the percentage on the left is higher than on the right. This, in turn, means that, on average, most queries with the baseline approach result in more relevant documents in the top 20 documents. If the figure on the left is much lower than the one on the right, then the baseline approach is worse than the given passage-based ranking. With this definition in mind, the general observation from Table 6 is that PAGES are consistently better than PARA-

TABLE 6. Comparison of precision at 20 documents of PAGES versus other passages.

	PARA	SEC	TILE	WIN-150	WIN-350
Short queries					
FR-12	**24/5**	**29/5**	29/10	19/14	19/10
FR-24	12/12	19/12	19/15	4/12	8/19
TREC-24	28/24	—	**40/12**	22/28	16/26
TREC-45	**38/8**	—	28/18	**42/20**	28/30
WSJ-12	34/30	—	30/32	**48/16**	28/30
Long queries					
FR-12	19/19	43/14	24/24	33/19	24/29
FR-24	12/8	**42/15**	**42/8**	19/15	19/12
TREC-24	**40/14**	—	34/22	**40/22**	42/18
TREC-45	28/26	—	20/20	44/38	38/38
WSJ-12	32/26	—	28/26	38/38	28/40

Each entry has two numbers, X and Y (that is, X/Y). X is the percentage of queries where PAGES order documents better than the given passage-ordering technique. Y is the percentage of queries where PAGES order documents worse than the given passage-ordering technique. The numbers in bold represent the significant results using the Wilcoxon test with 95% confidence level.

TABLE 7. Comparison of whole-document ranking with best passage ranking method.

	Precision at *N* documents						
	5	10	20	30	200	AvgP	%Δ
Short queries							
FR-12							
Document	0.1619	0.1429	0.1024	0.0937	0.0329	0.2075	0.0
PAGES	0.2095	0.1667	0.1238	0.1079	0.0400	0.3067	+47.8
FR-24							
Document	0.1615	0.1000	0.0750	0.0538	0.0148	0.1225	0.0
PARAGRAPHS	0.1692	0.1192	0.0788	0.0628	0.0173	0.1434	+17.1
TREC-24							
Document	0.3120	0.2840	0.2300	0.2060	0.0963	0.1348	0.0
PAGES	0.2920	0.2640	0.2130	0.1907	0.0916	0.1389	+3.0
TREC-45							
Document	0.4160	0.3520	0.3180	0.2740	0.1184	0.1909	0.0
PAGES	0.4320	0.3520	0.3070	0.2647	0.1159	0.1910	+0.1
WSJ-12							
Document	0.4160	0.4140	0.3700	0.3507	0.1963	0.2313	0.0
PAGES	0.4440	0.4120	0.3720	0.3493	0.2066	0.2450	+5.9
Long queries							
FR-12							
Document	0.1810	0.1571	0.1357	0.1365	0.0533	0.2417	0.0
PARAGRAPHS	0.3048	0.2571	0.2143	0.1905	0.0700	0.3616	+49.6
FR-24							
Document	0.1923	0.1385	0.1115	0.0923	0.0237	0.1543	0.0
WINDOWS-150	0.2231	0.1923	0.1404	0.1141	0.0283	0.2220	+43.9
TREC-24							
Document	0.4400	0.3860	0.3270	0.2847	0.1195	0.11883	0.0
PAGES	0.4240	0.4000	0.3310	0.2760	0.1247	0.1958	+4.0
TREC-45							
Document	0.5160	0.4240	0.3310	0.2880	0.1234	0.2037	0.0
PAGES	0.4920	0.4320	0.3320	0.2960	0.1324	0.2182	+7.1
WSJ-12							
Document	0.6320	0.5560	0.5030	0.4647	0.2607	0.3230	0.0
WINDOWS-150	0.6200	0.5760	0.5030	0.4620	0.2567	0.3283	+1.6

GRAPHS, SECTIONS, and TILES. The best case for other passages is when they are even with PAGES. However, despite the fact that there are queries when document retrieval based on passages succeeds, an improper segmentation that does not reflect the query can be detrimental.

The experimental results from this section are summarised in Table 7 where, for each of the 10 test environments, whole-document ranking is compared with the *best* passage-based ranking. Two points should be made. First, the most consistent method based on passages is PAGES. Out of 10 tests, PAGES performed best six times. For each of these, it was better than whole-document ranking, averaging an improvement of over 10% per test case. In other tests, PARAGRAPHS was slightly superior to other passages, and in one case WINDOWS-150 was best.

The second point is that, even though PAGES is the best performing passage-based method, there is room for further improvement. For example, even though it works well with short queries, the results for long queries are mixed. We conclude that this is due to poor segmentation of long documents. Case analysis of individual queries (Kaszkiel, 2000) revealed that a long relevant document can be greatly penalised if only a short fragment is relevant to the query. Passage-based ranking avoids this problem by estimating the document's relevance using only a single fragment.

An indirect result of our experiments is confirmation that pivoted document length normalization (Singhal et al., 1996a, 1998) is a successful innovation. For collections of text of varying length—in particular whole documents, sections, or paragraphs—it gave a marked improvement in effectiveness.

Arbitrary Passages

Passages of the types discussed in the previous section were defined before or during indexing, which has several consequences. First, documents are partitioned into passages without consideration of individual queries. Second, when discourse passages such as paragraphs are used, long sections may be split into passages that are individually less informative, which is undesirable if the entire section is relevant to a given query. Splitting a relevant passage into parts is referred to as *blurring* (Stanfill & Waltz, 1992). Third, the definition of a passage is subjective, and depends on document structure. For instance, assuming that discourse passages are used in a collection of journal articles,

TABLE 8. Improvements in retrieval effectiveness of any single predefined passage type compared with whole-document ranking (Predefined), and improvements in retrieval effectiveness of the best predefined passage type selected for each query compared with whole-document ranking (BestOfAll).

	Collection				
	FR-12	FR-24	TREC-24	TREC-45	WSJ-12
Short queries					
Predefined	PAGES	PARAGRAPHS	PAGES	PAGES	PAGES
%Δ	+47.8	+17.1	+3.0	+0.1	+5.9
BestOfAll (%Δ)	+49.3	+30.1	+10.3	+6.7	+14.1
Long queries					
Predefined	PARAGRAPHS	WINDOWS-150	PAGES	PAGES	WINDOWS-150
%Δ	+49.6	+43.9	+4.0	+7.1	+1.6
BestOfAll (%Δ)	+64.3	+71.5	+19.8	+17.8	+13.8

The improvements (%Δ) are for average precision (AvgP). The pivoted-cosine measure was used in all cases.

in some cases users might want to retrieve sections and, in others, paragraphs.

The effectiveness of previous passage types varied, and did not identify a clear winner. Also, it is not clear whether the limit of passage retrieval was reached. To explore the improvement that is potentially available, we determined the best possible retrieval result available using the passage types tested. The effectiveness associated with the best passage type for each query was selected, then these "bests" were averaged over the query set. The percentage improvement of the "best" result of predefined passage types compared with whole-document ranking is shown in Table 8. The results show that (not surprisingly) higher effectiveness is available if, post hoc, the best passage type is selected for each query. The improvements in average precision for short queries is not as significant as for long queries. However, for collections of uniform length such as TREC-24, TREC-45, and WSJ-12, even though passage ranking is not expected to affect retrieval significantly, consistent improvements in effectiveness demonstrate that passage retrieval can be valuable if the right passage types are selected.

Fixed-Length Arbitrary Passages

To explore whether better passage selection is possible, we propose an alternative to the passage types discussed in the previous section. We define an *arbitrary passage* as any sequence of words of any length starting at any word in the document. The locations and dimensions of passages are delayed until the query is evaluated, so that the similarity of the highest-ranked sequence of words, from anywhere in the document, defines the passage to be retrieved; or, in the case of document retrieval, determines the document's similarity. Two subclasses are defined, fixed-length passages, where the length of the passage is set before query evaluation, and variable-length passages, where passages can be of any length.

The definition of fixed-length arbitrary passages is similar to the *sliding window* used by Callan (1994), who defines the first sliding window in each document as starting at the first occurrence of a query term. Subsequent windows half-overlap preceding ones. Sliding window and fixed-length arbitrary passages are similar, but there is a distinctive difference: the number of possible passages in a document using sliding windows depends on passage length—the longer the sliding window, the smaller the number of passages. In contrast, fixed-length arbitrary passages can start at any word in the document.

Clarke et al. (1995) introduced a language that supports Boolean queries for any text *segment* in a collection, considered as the shortest unit of text that satisfies a Boolean query. This approach is not unlike using fixed-length passages, but is differentiated by the Boolean-based approach, which considers the importance of neither term nor document statistics. A similar approach by Hawking and Thistlewaite (1995) uses proximity of query terms to rank documents. A strength of both approaches is their applicability to distributed text collections, as both are independent of global statistics.

Melluci (1998) uses a probabilistic approach to extract passages. Bayesian statistics determine the degree to which query terms are concentrated more in relevant documents than irrelevant ones. The probabilistic approach requires enough information for the weight of terms to be estimated reliably, which in turn, leads to problems for passages, because generally passages are short and there is little consistency in the different term distributions. As a solution to this problem, Melluci uses a Bayesian framework to estimate the weights of terms in passages. These weights are calculated using the prior and current concentrations of terms in text. This approach has a more theoretical framework than fixed-length arbitrary passages, but incorporates many variables and is computationally expensive (Melucci, 1998).

Instead of defining passages, Mittendorf and Schäuble (1994) use inferred passage boundaries, by employing a hidden Markov model to determine passages appropriate to each query. This approach is analogous to TextTiling (1993), but passage boundaries are determined at query time

instead of indexing time. This approach necessitates processing of the full text to evaluate a query, but does demonstrate the ability of passage ranking to improve effectiveness.

Fixed-length arbitrary passages do have one serious drawback: naively implemented, the cost of ranking passages is high. The number of candidates for passages in a collection is much larger than the number for documents or predefined passages, and so ranking is more expensive, and impractical. However, our separate exploration of the issue of efficient passage ranking shows that it is practical on a desktop machine (Kaszkiel et al., 1999). With conventional ranking algorithms, passage ranking can be extremely costly; it is not feasible, for example, to use the strategy of allocating an accumulator to each unit to be ranked. We found that the costs of passage ranking can be greatly reduced by employing strategies to rapidly identify a small number of passages for which it is worthwhile computing a similarity; these are the passages containing the rarest query terms. By using efficient document-ordered merging of inverted lists for rare terms to choose passages, then using efficient term-ordered list intersections to complete the similarity computation, passages can be ranked in only a small multiple of the time required for document ranking.

Experiments with Fixed-Length Arbitrary Passages

In this section, we present results of experiments using fixed-length arbitrary passages for document ranking. To make the comparison between different passage lengths practicable, we used the following heuristics. We chose a set of fixed passage lengths from 50 to 600 words in increments of 50, that is, 12 different lengths. Passages of 600 words seemed a reasonable maximum as this figure well exceeds the median document length for the TREC data, while passages of less than 50 words are not likely to capture the information need. To limit the costs of query evaluation and to simplify implementation, passages start at 25-word intervals, which was earlier shown by us (Kaszkiel & Zobel, 1997) to be as effective as passages that start at every word in a document. Some less effective passage lengths are omitted from the tables in this section.

Experimental results have shown that the pivoted-cosine measure is superior when ranking units that vary in length, but with predefined passages such as WINDOWS, the cosine measure is as effective as the pivoted-cosine measure. As a consequence, we used the cosine measure to compute the similarities of fixed-length passages. The individual components of the inverse document frequency, f_t and N, were computed as if the database is a collection of documents. These variables could reflect the number of passages in the collection and the number of passages in which words occur. However, how to compute f_t and N to reflect fixed-length arbitrary passages instead of documents is not clear.

Our experiments investigate the effectiveness of ranking fixed-length arbitrary passages compared with whole-documents and predefined passage types such as PAGES, WINDOWS, PARAGRAPHS, and TILES. For each passage length, a single passage is used to estimate the document's similarity to a query. Whole-document ranking is calculated with the pivoted-cosine measure. To compare document retrieval based on fixed-length passages with predefined passage types, results for the best predefined passage type for each collection and query set are shown in each table. The results for predefined passages are calculated using the pivoted-cosine measure.

FR-12 collection

The retrieval effectiveness for fixed-length arbitrary passages, whole-document ranking, and the best predefined passage type, is shown in Table 9. The column marked as "%Δ" represents the change in the average precision from the baseline run, which is whole-document ranking using the pivoted-cosine measure.

For short queries, fixed-length arbitrary passages are better than whole-document ranking by up to 40%. However, fixed-length arbitrary passages are not as effective as the best predefined passage type, in this case PAGES. For long queries, similar trends are observed. The consistent effectiveness for different passage lengths is quite remarkable. For both query sets, any passage length in the range of 50–450 words outperforms whole-document ranking. Both methods—PAGES and fixed-length passages—are far superior to whole-document ranking.

Comparing the FR-12 and FR-24 collections, the effectiveness of predefined passage types varied, depending on query types and test collections. In one case PAGES was best, and in another PARAGRAPHS was best. Consistent improvements over whole-document ranking were achieved using fixed-length arbitrary passages. Irrespective of the query length, the worst case was when document retrieval using fixed-length passages was only as good as document retrieval using the best predefined passage technique. However, for most fixed-length passages, the retrieval was better than that achieved by PAGES.

TREC-45 collection

The results achieved by fixed-length arbitrary passage ranking on the FR-12 collection is promising. However, for large text collections with documents of more uniform length than those in the FR collections, whole-document ranking with the pivoted-cosine measure is expected to perform reasonably well (Kasziel & Zobel, 1997; Singhal et al., 1996a), thus reducing the benefits of passage retrieval.

Results for the TREC-45 collection are summarised in Table 10. There is a marked improvement when using fixed-length arbitrary passages to rank documents, especially for long queries. Compared with the best predefined passage, in this case PAGES, there is an improvement up to 4.2% using short queries and up to 14.1% using long queries. The results with short queries show the same trends as those observed with the FR-24 and TREC-24 results, where document ranking based on fixed-length passages is more

TABLE 9. FR-12 collection: document retrieval using fixed-length arbitrary passages.

	Precision at *N* documents						
	5	10	20	30	200	AvgP	%Δ
Short queries							
Document	0.1619	0.1429	0.1024	0.0937	0.0329	0.2075	0.0
PAGES	0.2095	0.1667	0.1238	0.1079	0.0400	0.3067	+47.8
Fixed-length arbitrary passages							
50 words	0.2000	0.1333	0.1095	0.1032	0.0395	0.2155	+3.9
100 words	0.1810	0.1476	0.1190	0.1143	0.0402	0.2793	+34.6
150 words	0.1905	0.1476	0.1262	0.1159	0.0414	0.2762	+33.1
200 words	0.1714	0.1571	0.1310	0.1159	0.0419	0.2747	+32.4
250 words	0.1714	0.1429	0.1357	0.1190	0.0438	0.2808	+35.3
300 words	0.1905	0.1524	0.1286	0.1159	0.0431	0.2813	+35.6
350 words	0.1905	0.1476	0.1238	0.1111	0.0445	0.2912	+40.3
400 words	0.1905	0.1524	0.1262	0.1127	0.0438	0.2817	+35.8
450 words	0.1810	0.1476	0.1214	0.1127	0.0440	0.2523	+21.6
Long queries							
Document	0.1810	0.1571	0.1357	0.1365	0.0533	0.2417	0.0
PARAGRAPHS	0.3048	0.2571	0.2143	0.1905	0.0700	0.3616	+49.6
Fixed-length arbitrary passages							
50 words	0.3048	0.2476	0.1976	0.1603	0.0731	0.2815	+16.5
100 words	0.3143	0.2429	0.2048	0.1778	0.0726	0.3343	+38.3
150 words	0.3048	0.2476	0.2071	0.1810	0.0736	0.3540	+46.5
200 words	0.3238	0.2762	0.2190	0.1889	0.0750	0.3252	+34.5
250 words	0.3238	0.2667	0.2190	0.1905	0.0752	0.3298	+36.5
300 words	0.3048	0.2762	0.2167	0.1968	0.0748	0.3257	+34.8
350 words	0.3238	0.2667	0.2143	0.1937	0.0736	0.3449	+42.7
400 words	0.2952	0.2667	0.2119	0.1873	0.0724	0.3405	+40.9
450 words	0.3143	0.2524	0.2167	0.1905	0.0731	0.3449	+42.7

effective than either whole-document ranking or most predefined passage types. However, fixed-length passages are 14.1% better than PAGES and 22.2% better than whole-document ranking. Even though fixed-length passages are more robust than whole-document ranking, the improvements for the TREC collections are not as large as for the FR collections.

Overall, with documents of uniform length, document retrieval using whole-document ranking with pivoted-cosine measure is almost as effective as with fixed-length arbitrary passage ranking. However, consistent improvements, sometimes small, are achieved with fixed-length arbitrary passages. The experiments using fixed-length passage ranking confirm the hypothesis that ordering documents according to a single segment is at least as effective as considering the entire document, and that document retrieval using fixed-length passages improves consistently over that for predefined passage types. The retrieval effectiveness is also consistent for different passage lengths. This shows the robustness of fixed-length passage ranking.

WSJ-12 collection

In general, passages should not have an impact on the effectiveness of document retrieval when most documents are short. The majority of documents in the WSJ-12 collection are shorter than 400 words. The best predefined passage type was PAGES for short queries and WINDOWS-150 for long queries. The best predefined passage type for the short query set, PAGES, showed most significant improvement compared with whole-document ranking, a 5.9% increase in average precision. For long queries, the difference between whole-document ranking and predefined passage-based ranking was mild. However, use of fixed-length arbitrary passages yielded small further improvements in effectiveness; results are not shown.

Significance and Analysis

The effectiveness of fixed-length arbitrary passages is not particularly sensitive to passage length, for lengths close to that achieving the best retrieval effectiveness. For example, for TREC-45 with long queries, average precision results (AvgP) for passage lengths of 50, 150, 200, and 250 are quite similar to the results for the passage length of 100 words, which performs best on average.

This result is confirmed in other work. A study by Allan (1995) showed that relevance feedback that uses passages instead of documents improves retrieval, with the best results achieved with passages of 200–300 words. In the context of document retrieval, our results confirm this, because the average best passage would be between 100 and 400 words, for short and long queries respectively. Papka

TABLE 10. TREC-45 collection: document retrieval using fixed-length arbitrary passages.

	Precision at *N* documents						
	5	10	20	30	200	AvgP	%Δ
Short queries							
Document	0.4160	0.3520	0.3180	0.2740	0.1184	0.1909	0.0
PAGES	0.4320	0.3520	0.3070	0.2647	0.1159	0.1910	+0.1
Fixed-length arbitrary passages							
50 words	0.3640	0.3000	0.2630	0.2320	0.1136	0.1835	−3.9
100 words	0.3800	0.3380	0.2850	0.2500	0.1149	0.1924	+0.8
150 words	0.3680	0.3420	0.2820	0.2467	0.1202	0.1939	+1.6
200 words	0.4000	0.3440	0.2880	0.2567	0.1232	0.1960	+2.7
250 words	0.4040	0.3540	0.2970	0.2573	0.1221	0.1966	+3.0
300 words	0.4040	0.3640	0.3020	0.2640	0.1221	0.1991	+4.3
350 words	0.4040	0.3480	0.3010	0.2660	0.1210	0.1958	+2.6
400 words	0.4120	0.3420	0.3040	0.2680	0.1205	0.1954	+2.4
450 words	0.3920	0.3460	0.3060	0.2667	0.1192	0.1966	+3.0
Long queries							
Document	0.5160	0.4240	0.3310	0.2880	0.1234	0.2037	0.0
PAGES	0.4920	0.4320	0.3320	0.2960	0.1324	0.2182	+7.1
Fixed-length arbitrary passages							
50 words	0.4720	0.3940	0.3270	0.2900	0.1313	0.2356	+15.7
100 words	0.5080	0.4180	0.3540	0.3120	0.1347	0.2489	+22.2
150 words	0.4640	0.4060	0.3580	0.3107	0.1386	0.2366	+16.2
200 words	0.4400	0.4100	0.3410	0.2973	0.1357	0.2270	+11.4
250 words	0.4720	0.4120	0.3460	0.3033	0.1347	0.2265	+11.2
300 words	0.4760	0.4080	0.3520	0.3033	0.1357	0.2255	+10.7
350 words	0.4840	0.4180	0.3430	0.2993	0.1334	0.2182	+7.1
400 words	0.4640	0.4120	0.3350	0.2967	0.1300	0.2118	+4.0
450 words	0.4560	0.3980	0.3320	0.2853	0.1261	0.2057	+1.0

and Allan (1997) used windows of text, which can be considered as passages, for massive query expansion, an automatic relevance feedback method that aims to add hundreds of new words to the original query. Their experimental results with a subset of the TREC data showed that longer passages give the best effectiveness. This confirms our results with short queries, where short passages provide too little context to make any judgments of documents or other relevant terms.

Our results and the work reported by others lead us to the following recommendations: for long queries, on average at least 10 words, best results are achieved with passages of 100 to 200 words; and for short queries, which are usually no more than three words, best results are achieved with passages of 250 to 350 words.

Whether document retrieval based on fixed-length passages significantly improves over whole-document ranking or predefined passage-based ranking is not clear. We compare document retrieval based on three ranking techniques: whole-document ranking, predefined passage ranking using PAGES, and fixed-length arbitrary passage ranking, with the emphasis on the difference from fixed-length arbitrary passage ranking. Table 11 summarizes two distinct results. The first result is the count of queries where there is a difference between two retrieval techniques. This is expressed in terms of the proportion of queries that differ. The second result is the test for the statistical significance; significant differences are shown in bold.

For short queries, the distinction between fixed-length arbitrary passages and other techniques is not clear. There is a significant difference between document retrieval using fixed-length passages and whole documents for only two test collections, FR-12 and WSJ-12. In terms of the number of queries with different average precision, there is no difference between fixed-length passage ranking and predefined passage ranking such as PAGES. For long queries, the difference between document retrieval using fixed-length

TABLE 11. Comparison of retrieval effectiveness (AvgP) of fixed-length arbitrary passages with whole-document ranking and PAGES-based ranking.

	Short queries		Long queries	
	Document	PAGES	Document	PAGES
FR-12	**52/29**	33/43	67/24	43/38
FR-24	42/42	46/31	**65/31**	**65/27**
TREC-24	38/56	46/48	60/36	58/38
TREC-45	46/50	54/44	**60/40**	50/50
WSJ-12	**64/36**	56/44	58/42	58/42

Each entry has two numbers, *X* and *Y* (that is, *X*/*Y*). *X* is the percentage of queries where the given fixed-length passage ranking is better than whole-document ranking or PAGES-based ranking. *Y* is the percentage of queries where the given fixed-length passage ranking is worse than whole-document ranking or PAGES-based ranking. The numbers in bold represent the statistically significant results using the Wilcoxon test with a 95% confidence level. Recommended passage lengths are used: 250 words for short queries and 150 words for long queries.

passage ranking and the other two techniques is more distinct. In all cases but one, the TREC-24 collection with predefined passages, document retrieval based on fixed-length passages produces more queries with higher average precision than whole-document ranking or predefined passage ranking.

In summary, in the five test collections, retrieval based on fixed-length arbitrary passages was found to be significantly better than document ranking, for both short and long queries. However, comparing document retrieval based on fixed-length passages and predefined passage such as PAGES, no significant differences were found, except on the FR-24 and TREC-45 collections.

TABLE 12. Improvements in retrieval effectiveness of any single fixed-length passage compared with whole-document ranking (Fixed), and improvements in retrieval effectiveness of the best fixed-length passage selected for each query compared with whole-document ranking (BestFixed).

	Collection				
	FR-12	FR-24	TREC-24	TREC-45	WSJ-12
Short queries					
Fixed (%Δ)	+40.3	+21.1	+2.4	+4.3	+7.0
BestFixed (%Δ)	+52.9	+45.6	+16.0	+13.7	+17.1
Long queries					
Fixed (%Δ)	+46.5	+50.2	+10.6	+22.2	+5.7
BestFixed (%Δ)	+65.9	+73.5	+19.8	+36.2	+17.3

The improvements (%Δ) are for average precision (AvgP).

Variable-Length Arbitrary Passage Retrieval

Our results show that, on average, document retrieval using fixed-length passages is at least as effective as with predefined passages, which we have also reported elsewhere (Kaszkiel & Zobel, 1997). The studies of retrieval results for individual queries showed that no particular length was superior. That is, for queries of the same type, one passage length worked best for some queries but not for others. A solution to the limitations of fixed-length arbitrary passages is to select a passage length most likely to suit the query. The best passage length can also depend on the documents ranked. For example, given a query, we could find two long documents, where in one the start of document or the abstract is relevant, and in the other a 400-word section is relevant. Adjusting the passage length to the type of text should result in improved retrieval.

Therefore, a more flexible approach would be to extract passages of different lengths, and select the best one to represent each document. We refer to this approach as *variable-length arbitrary passage* retrieval. A variable-length passage is of any length that is determined by the best passage in a document, when the query is evaluated. Our earlier preliminary studies were encouraging (Kaszkiel & Zobel, 1997).

Assuming that for each query average precision is that of the best fixed-length passage type, the retrieval effectiveness is expected to be higher than that for the best result with all predefined passage types (see Table 8). The improvements in average precision when the best passage length is chosen for each query is shown in Table 12. The improvements are consistently higher than those for the best predefined passages, in particular for short queries. These results indicate that further improvements are possible if passage length is varied.

Variable-length arbitrary passage ranking is similar to locality based retrieval, as proposed by de Kretser and Moffat (1999), where document boundaries are ignored and text is treated as a continuous sequence of words. The similarity scores for passages are according to how many query term occurrences appear near to each other. Shape, height, and spread of a function is used to calculate the contribution of query terms to text regions. High-scoring regions are identified and passages that contain them retrieved. In this approach, the length of the passages depends on a scoring function and the corresponding parameters are used to identify text regions. The parameters used in the function need to be adjusted for different collections and query sets. No consistent results for any functions were reported (de Kretser & Moffat, 1999).

In fixed-length passage retrieval we calculate the similarity of each passage independent of its length. Thus, documents are ranked according to the best passage from each document. However, for variable-length passage ranking, documents are represented by passages of different lengths, so there are two related problems: first, how to discriminate between passages of different length in the same document; second, how to discriminate between passages of different lengths drawn from different documents.

In the absence of length normalisation in the similarity measure, the longest passage for each document determines the rank of the document. This is undesirable because, as we found for fixed-length passages, effectiveness degrades with passages in excess of 450 words. To select a passage to represent a document, pivoted-cosine normalization can be used, which is restated here for variable-length passages:

$$W_p = (1 - \text{slope}) + \text{slope} \cdot \frac{p_{len}}{\Delta_{len}}$$

where *slope* is set to 0.2 (which was shown to be effective in the context of predefined passage ranking [16]), p_{len} is the length of fragment p in bytes, and Δ_{len} is the average length of all fragments in the collection. This formula has been shown to be effective for predefined passage types and minimizes the fragility of ranking fragments of varying length. The overall similarity of passage p to a query q is:

$$\frac{\text{sim}(q, p)}{W_p}.$$

Formally, this is not applicable to variable-length passage ranking because it requires averages over all passage

lengths in the collection, which is not meaningful in the context of variable-length passages. Singhal et al. (1996a) have argued that this length formulation is reasonably robust if Δ_{len} is set to an overall average, which in this case is the average passage length used (about 300 words). This approach is referred to as *Variable.* The similarity score for a document d to a query q is based on the best-scoring passage among 12 different lengths in the range of 50 to 600 words:

$$\text{sim}(q, d) = \max\left(\frac{\text{sim}(q, d, p_{50})}{W_{p,50}}, \frac{\text{sim}(q, d, p_{100})}{W_{p,100}}, \ldots, \frac{\text{sim}(q, d, p_{600})}{W_{p,600}}\right)$$

where $\text{sim}(q, d, p_{len})$ is the similarity of passage p of length len in document d to query q, based on the cosine measure. The value of $W_{p,len}$ is the pivoted-cosine normalisation component for passage p of length len.

Experiments with Variable-Length Arbitrary Passages

In these experiments, we restricted the passage lengths to the set 50, 100, 150, . . . , 600 words, which were used for experiments with fixed-length passages. For evaluation to be consistent with previous experiments, only a single passage of any length is used as document evidence.

Table 13 shows results for the *Variable* strategy for variable-length passage ranking. The change in average precision, or %Δ, is measured against the baseline result of whole-document ranking using the pivoted-cosine measure. For the *Variable* approach, we determined (experimentally on one test collection) that the best document ranking is achieved when Δ_{len} is set to around the best fixed-length passage for a particular type of queries. For long queries, best results are achieved with Δ_{len} set to 100 words and, for short queries, set to 300 words. All possible passage lengths between 50 and 600 words are used. A range of slopes was experimented with; the most consistent was 0.2 for whole-document and predefined passage ranking. Thus, *slope* is set to 0.2.

For short queries, the *Variable* strategy achieves the best average precision across the five collections. The improvements over the baseline range from 5.8% for TREC-45 to 42.6% for FR-12. In addition, for collections with many long documents such as FR-12 and FR-24, the precision at the 5 and 10 document cutoffs is significantly higher than for whole-document ranking. For both evaluation measures, average precision and precision at low document cutoffs, the *Variable* approach is an improvement on the average effectiveness of the recommended fixed-length passage. Even selecting the best passage for each collection does not perform as well as *Variable.* These results support the supposition that no single passage length suits the matching between queries and documents.

For long queries, we observe similar trends. The relative improvement in average precision (%Δ) for each collection is larger than for short queries, ranging from 10.7 to 50.6%. Similarly, the difference in precision at document 5 and 10 cutoffs is higher with variable-length passages than with either whole-document or fixed-length passage ranking.

TABLE 13. Retrieval results for variable-length arbitrary passages.

	Precision at *N* documents						
	5	10	20	30	200	AvgP	%Δ
Short queries							
FR-12							
Document	0.1619	0.1429	0.1024	0.0937	0.0329	0.2075	0.0
Variable	0.1905	0.1619	0.1286	0.1175	0.0424	0.2960	+42.6
FR-24							
Document	0.1615	0.1000	0.0750	0.0538	0.0148	0.1225	0.0
Variable	0.1769	0.1269	0.0904	0.0679	0.0163	0.1548	+26.4
TREC-24							
Document	0.3120	0.2840	0.2300	0.2060	0.0963	0.1348	0.0
Variable	0.3040	0.2820	0.2280	0.1920	0.0942	0.1444	+7.1
TREC-45							
Document	0.4160	0.3520	0.3180	0.2740	0.1184	0.1909	0.0
Variable	0.4080	0.3440	0.3020	0.2673	0.1224	0.2020	+5.8
WSJ-12							
Document	0.4160	0.4140	0.3700	0.3507	0.1963	0.2313	0.0
Variable	0.4200	0.4000	0.3730	0.3527	0.2143	0.2532	+9.5
Long queries							
FR-12							
Document	0.1810	0.1571	0.1357	0.1365	0.0533	0.2417	0.0
Variable	0.3143	0.2524	0.2238	0.1810	0.0745	0.3611	+49.4
FR-24							
Document	0.1923	0.1385	0.1115	0.0923	0.0237	0.1543	0.0
Variable	0.1846	0.1808	0.1462	0.1192	0.0306	0.2323	+50.6
TREC-24							
Document	0.4400	0.3860	0.3270	0.2847	0.1195	0.1883	0.0
Variable	0.4520	0.3920	0.3120	0.2747	0.1269	0.2104	+11.7
TREC-45							
Document	0.5160	0.4240	0.3310	0.2880	0.1234	0.2037	0.0
Variable	0.4800	0.4400	0.3600	0.3240	0.1440	0.2574	+26.4
WSJ-12							
Document	0.6320	0.5560	0.5030	0.4647	0.2607	0.3230	0.0
Variable	0.6320	0.5540	0.5220	0.4873	0.2806	0.3577	+10.7

The improvements in average precision (%Δ) are over whole-document ranking with pivoted-cosine measure.

Significance and Analysis

The comparison of variable-length arbitrary passage ranking to whole-document ranking suggests that passages are more effective at retrieving relevant documents. In this section we compare variable-length passage ranking with recommended fixed-length arbitrary passage ranking. The results for both, fixed-length and variable-length passages, are shown in Table 14.

For short queries, variable-length passages consistently improve retrieval compared with fixed-length passages. However, the gains are not significant; they range from 1.4 to 5.4% over that for the recommended passage length of 250 words. For long queries, the effectiveness of variable-length and fixed-length arbitrary passages on FR-12, FR-24, and TREC-24 is comparable, but slightly in favor of vari-

TABLE 14. Comparison of document retrieval with variable-length arbitrary passages and recommended fixed-length arbitrary passages.

	Precision at *N* documents						
	5	10	20	30	200	AvgP	%Δ
Short queries							
FR-12							
250 words	0.1714	0.1429	0.1357	0.1190	0.0438	0.2808	0.0
Variable	0.1905	0.1619	0.1286	0.1175	0.0424	0.2960	+5.4
FR-24							
250 words	0.1692	0.1269	0.0865	0.0679	0.0171	0.1527	0.0
Variable	0.1769	0.1269	0.0904	0.0679	0.0163	0.1548	+1.4
TREC-24							
250 words	0.3040	0.2800	0.2190	0.1933	0.0936	0.1380	0.0
Variable	0.3040	0.2820	0.2280	0.1920	0.0942	0.1444	+4.6
TREC-45							
250 words	0.4040	0.3540	0.2970	0.2573	0.1221	0.1966	0.0
Variable	0.4080	0.3440	0.3020	0.2673	0.1224	0.2020	+2.7
WSJ-12							
250 words	0.4520	0.4040	0.3710	0.3460	0.2097	0.2470	0.0
Variable	0.4200	0.4000	0.3730	0.3527	0.2143	0.2532	+2.5
Long queries							
FR-12							
150 words	0.3048	0.2476	0.2071	0.1810	0.0736	0.3540	0.0
Variable	0.3143	0.2524	0.2238	0.1810	0.0745	0.3611	+2.0
FR-24							
150 words	0.2000	0.1808	0.1327	0.1167	0.0294	0.2296	0.0
Variable	0.1846	0.1808	0.1462	0.1192	0.0306	0.2323	+1.2
TREC-24							
150 words	0.4640	0.3860	0.3200	0.2753	0.1255	0.2082	0.0
Variable	0.4520	0.3920	0.3120	0.2747	0.1269	0.2104	+1.1
TREC-45							
150 words	0.4640	0.4060	0.3580	0.3107	0.1386	0.2366	0.0
Variable	0.4800	0.4400	0.3600	0.3240	0.1440	0.2574	+8.8
WSJ-12							
150 words	0.5520	0.5380	0.4920	0.4693	0.2705	0.3409	0.0
Variable	0.6320	0.5540	0.5220	0.4873	0.2806	0.3577	+4.9

able-length passages. However, for the TREC-45 and WSJ-12 collections, the retrieval improvements for variable-length passages are up to 8.8%. In conclusion, the additional gains from using variable-length passages are not as high as expected.

We use the Wilcoxon signed rank test with 95% confidence level to discover any statistically significant differences in retrieval effectiveness. We compare document retrieval using three techniques: variable-length passages with *Variable* normalization, whole-document, and PAGES ranking. Results are shown in Table 15.

The improvement in average precision for individual queries is not as evident for short queries as for long queries. The only significant difference is on the FR-12 and WSJ-12 collections, where variable-length passages improve over whole-document ranking and PAGES ranking. This result confirms the significance of improvements with fixed-length passages shown in Table 11. For long queries, the evidence for significant improvements of variable-length arbitrary passages compared with whole-document ranking and predefined passages is stronger. For all collections, document retrieval using variable-length passage ranking as opposed to whole-document ranking improves effectiveness for most queries. Furthermore, for all collections except FR-24, the improvement is statistically significant. This is in contrast to fixed-length passage ranking (Table 11), where results on only two collections were statistically significant.

Compared with the best predefined passage ranking, the effectiveness of variable-length arbitrary passage ranking is consistently improved. For all collections, the majority of queries are better with variable-length passages than with PAGES. In addition, 7 out of 10 results are significant in favor of variable-length passages.

Comparison of variable-length arbitrary passages with the best predefined passage type shows that they consistently perform better. However, one of the aims of variable-length passages is to achieve effectiveness similar to that of a system that can select the best predefined passage type for each query. To investigate whether this is the case, we compare variable-length passage-based ranking with a system that chooses the best predefined passage type for each query. The comparison is shown in Table 16, which indicates that variable-length passage-based ranking does not achieve the same effectiveness as a system that can select the best predefined passage type for each query. For short queries most differences are significant, despite the fact that the absolute difference in precision improvements over whole-document ranking is less than 7% for all collections (compare Tables 8 and 13). For long queries, for the majority of collections the differences are not statistically significant.

We believe that further improvements for variable-length passage-based ranking are possible if passage length normalization is refined to better discriminate between passages of varying length. We showed results for a system that could select the best fixed-length passage for each query (BestFixed) in Table 12. For all test cases, that is, varying collection and varying query length, the BestFixed approach is better than a system that can select the best predefined passage for each query (Table 8).

TABLE 15. Comparing the average precision (AvgP) of variable-length arbitrary passage ranking with whole-document ranking and document retrieval based on predefined passages, PAGES.

	Short queries		Long queries	
	Document	PAGES	Document	PAGES
FR-12	**57/24**	33/48	**67/24**	52/29
FR-24	50/38	42/38	58/38	58/35
TREC-24	42/48	50/46	**62/32**	**60/36**
TREC-45	48/50	54/44	**68/32**	**58/42**
WSJ-12	**68/30**	**64/34**	**68/32**	**72/38**

Each entry has two numbers, *X* and *Y* (that is, *X*/*Y*). *X* is the percentage of queries where the variable-length passage ranking technique is better than whole-document ranking or PAGES ranking. *Y* is the percentage of queries where the variable-length passage ranking technique is worse than whole-document ranking or PAGES ranking. The numbers in bold represent the significant results using the Wilcoxon test with a 95% confidence level.

TABLE 16. Comparing the average precision (AvgP) of variable-length arbitrary passage ranking (Pivoted approach) with retrieval based on selecting the best predefined passage type for each query.

	Query type	
	Short	Long
FR-12	24/52	**24/48**
FR-24	**31/58**	31/62
TREC-24	**36/60**	**38/58**
TREC-45	**34/62**	48/52
WSJ-12	**40/56**	42/58

Each entry has two numbers X and Y (that is, X/Y). X is the percentage of queries where the variable-length passage ranking technique is better than for the highest possible retrieval with all predefined passage types. Y is the percentage of queries where the variable-length passage ranking technique is worse than for the highest possible retrieval with all predefined passage types. The numbers in bold represent the significant results using the Wilcoxon test with a 95% confidence level.

Conclusions

Previous work has shown that document retrieval based on passage-based ranking is a promising approach. However, there has been no direct comparison of the effect of using different types of passages. We reviewed and evaluated passages based on discourse properties (PARAGRAPHS and SECTIONS), topical content (TILES), and nonoverlapping windows (WINDOWS and PAGES), all of which were the subjects of earlier research. We showed that these predefined passage types are generally more effective than whole documents at identifying relevant documents, in particular on text collections of varying document lengths or with many long relevant documents. The improvement obtained by passage ranking compared with whole-document ranking varied depending on the passage type, collection, and query set, with the greatest improvements in average precision for passage ranking ranging from 20 to 50%. For text collections with uniform document lengths, the improvements did not exceed 7%.

Despite the general improvements in effectiveness of passage-based ranking, no single passage type showed superior retrieval effectiveness across five different text collections and two query sets. To extend our studies into passages and their applications, we proposed arbitrary passages. Document retrieval with fixed-length arbitrary passages was shown to be more effective than with either whole-document ranking or predefined passage-based ranking. Retrieval via fixed-length passages consistently performs well across different collections and query sets. Per-query analysis showed that fixed-length passage ranking was significantly more effective than whole-document ranking but, except in two cases, no significant differences were found when compared with the best predefined passage-based ranking. Moreover, our experimental results showed that there is no single passage length that gives best effectiveness across the various collections and query sets; we found that, for short queries, longer passages of 250 and 350 words work best, while, for queries in excess of 10 words, the best results are achieved with shorter passages of 100 to 200 words. For short queries, the likelihood of finding query terms is higher in long passages than in short passages. For long queries, query terms are more likely to occur in close proximity; therefore, it is more important to locate short text segments that contain dense occurrences of query terms.

Document retrieval using variable-length arbitrary passages provided small further improvements in retrieval effectiveness compared with fixed-length arbitrary passage ranking. For long queries, the improvements were statistically significant for most collections, when compared with whole-document ranking and PAGES-based ranking. This is in contrast to fixed-length passage ranking where improvements on only two collections were found to be significant. Our objective in testing variable-length passages was to achieve a similar level of effectiveness to that achieved by selecting the best predefined passage for each query. Variable-length passages almost achieved this goal, but our results also showed that significant further gains may be possible.

The use of arbitrary passages in this article was limited to only one application: retrieving documents according to a single best passage. Possible extensions include document retrieval according to several highly ranked passages and passage-based query refinement, also known as automatic relevance feedback. Another application we are currently exploring is to use arbitrary passage retrieval for question answering. The aim is to apply information retrieval techniques, possibly in combination with natural language processing, to reduce the amount of text presented to users who require answers to specific questions.

Passages are an effective mechanism for information retrieval in environments in which other retrieval techniques can be poor: databases of long documents, of heterogeneous documents, and of data in which there are no predefined divisions into documents. In even standard collections of text, passages have the potential to improve effectiveness, and they help to locate relevant parts of documents. Their major potential drawback is the cost of query evaluation, but we have shown elsewhere that evaluation is feasible on a conventional machine (Kaszkiel, 2000; Kaszkiel et al., 1999). Passages are a method of choice for information retrieval.

Acknowledgment

This work was supported by the Australian Research Council.

References

Allan, J. (1995). Relevance feedback with too much data. In E.A. Fox, P. Ingwersen, & R. Fidel (Eds.), Proceedings of the 18th annual international ACM-SIGIR conference on research and development in information retrieval, Seattle, WA, July (pp. 337–343).

Beeferman, D., Berger, A., & Lafferty, J. (1997). Text segmentation using exponential models. In Proceedings of empirical methods in natural language processing 2, Providence, RI: AAAI Press.

Brandow, R., Mitze, K., & Rau, L. (1995). Automatic condensation of electronic publications by sentence selection. Information Processing and Management, 31(5), 675–685.

Buckley, C., Salton, G., Allan, J., & Singhal, A.K. (1994). Automatic query expansion using SMART: TREC 3. In D.K. Harman (Ed.), Proceedings of the 3rd text retrieval conference (TREC-3), NIST Special Publication 500–225 (pp. 69–80), Gaithersburg, MD: NIST.

Callan, J.P. (1994). Passage-retrieval evidence in document retrieval. In B.W. Croft & C.J. van Rijsbergen (Eds.), Proceedings of the 17th annual international ACM-SIGIR conference on research and development in information retrieval, Dublin, Ireland, July (pp. 302–310), New York: ACM.

Clarke, C., Cormack, G., & Burkowski, F. (1995). Shortest substring ranking (MultiText experiments for TREC-4). In D.K. Harman (Ed.), Proceedings of the 4th text retrieval conference (TREC-4), NIST special publication 500–236 (pp. 295–304), Gaithersburg, MD: NIST.

Cormack, G., Clarke, C., Palmer, C., & To, S. (1997). Passage-based refinement (MultiText experiments for TREC-6). In E.M. Voorhees & D.K. Harman (Eds.), Proceedings of the 6th text retrieval conference (TREC-6), NIST special publication 500–240 (pp. 303–320), Gaithersburg, MD: NIST.

Cormack, G., Palmer, C., Biesbrouck, M., & Clarke, C. (1998). Deriving very short queries for high precision and recall (MultiText experiments for TREC-7). In E.M. Voorhees & D.K. Harman (Eds.), Proceedings of the 7th text retrieval conference (TREC-7), NIST special publication 500–242 (pp. 121–132), Gaithersburg, MD: NIST.

Crestani, F., Lalmas, M., van Rijsbergen, C., & Campbell, I. (1998). A survey of probabilistic models in information retrieval. ACM Computing Surveys, 30(4), 528–552.

de Kretser, O., & Moffat, A. (1999). Locality-based information retrieval. In Proceedings of the 10th Australasian database conference, Auckland, New Zealand (pp. 177–188), Singapore: Springer-Verlag.

Harman, D. (1995). Overview of the second text retrieval conference (TREC-2). Information Processing and Management, 31(3), 271–289.

Hawking, D., & Thistlewaite, P. (1995). Proximity operators—So near and yet so far. In D.K. Harman (Ed.), Proceedings of the 4th text retrieval conference (TREC-4), NIST special publication 500–236 (pp. 131–143), Gaithersburg, MD: NIST.

Hearst, M. (1994). Multi-paragraph segmentation of expository texts. In Proceedings of the 32nd annual meeting of the association for computational linguistics, (pp. 9–16), Las Cruces, NM: ACL.

Hearst, M.A., & Plaunt, C. (1993). Subtopic structuring for full-length document access. In R. Korfhage, E. Rasmussen, & P. Willet (Eds.), Proceedings of the 16th annual international ACM-SIGIR conference on research and development in information retrieval, Pittsburgh, PA (pp. 59–68), New York: ACM.

Kaszkiel, M. (2000). Indexing and retrieval of passages in full-text databases. PhD thesis, Department of Computer Science.

Kaszkiel, M., & Zobel, J. (1997). Passage retrieval revisited. In N.J. Belkin, D. Narasimhalu, & P. Willett (Eds.), Proceedings of the 20th annual international ACM-SIGIR conference on research and development in information retrieval, Philadelphia, PA (pp. 178–185).

Kaszkiel, M., Zobel, J., & Sacks-Davis, R. (1999). Efficient passage ranking for document databases. ACM Transactions on Information Systems, 17(4), 406–439.

Keen, E.M. (1992). Presenting results of experimental retrieval comparisons. Information Processing and Management, 28(4), 491–502.

Lalmas, M., & Ruthven, I. (1997). A model for structured document retrieval: Empirical investigations. In N. Fuhr, G. Dittrich, & K. Tochtermann (Eds.), Hypertext—Information Retrieval—Multimedia (pp. 53–66). Dortmund, Germany.

Lu, X.A., & Keefer, R.B. (1994). Query expansion/reduction on its impact on retrieval effectiveness. In D.K. Harman (Ed.), Proceedings of the 3rd text retrieval conference (TREC-3), NIST special publication 500–225 (pp. 231–239), Gaithersburg, MD: NIST.

Maarek, Y.S., & Wecker, A.J. (1994). The librarian's assistant: Automatically organizing online books into dynamic bookshelves. In Proceedings of the international conference on intelligent multimedia information retrieval systems and management (pp. 233–247).

Melucci, M. (1998). Passage retrieval: A probabilistic technique. Information Processing and Management, 34(1), 43–68.

Mittendorf, E., & Schäuble, P. (1994). Document and passage retrieval based on hidden Markov models. In B.W. Croft & C.J. van Rijsbergen (Eds.), Proceedings of the 17th annual international ACM-SIGIR conference on research and development in information retrieval, Dublin-Ireland (pp. 318–327), New York: ACM.

Paice, C.D., & Jones, P.A. (1993). The identification of important concepts in highly structured technical papers. In R. Korfhage, E. Rasmussen, & P. Willet (Eds.), Proceedings of the 16th annual international ACM-SIGIR conference on research and development in information retrieval, Pittsburgh, PA (pp. 69–75), New York: ACM.

Papka, R., & Allan, J. (1997). Why bigger windows are better than smaller ones. Technical Report TR97-64, Department of Computer Science. Amherst, MA: University of Massachusetts.

Ponte, J.M., & Croft, B.W. (1997). Text segmentation by topic. In Proceedings of the 1st European conference on research and advanced technology for digital libraries (pp. 113–125).

Reynar, J.C. (1994). An automatic method of finding topic boundaries. In Proceedings of the 32nd annual meeting of the association for computational linguistics (student session), Las Cruces, NM: ACL.

Richmond, K., Smith, A., & Amitay, E. (1997). Detecting subject boundaries within text: A language independent statistical approach. In Proceedings of the 2nd conference on empirical methods in natural language processing, Providence, RI (pp. 47–54).

Robertson, S.E., & Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weight retrieval. In B.W. Croft & C.J. van Rijsbergen (Eds.), Proceedings of the 17th annual international ACM-SIGIR conference on research and development in information retrieval, Dublin, Ireland (pp. 232–241), New York: ACM.

Robertson, S.E., Walker, S., & Beaulieu, M. (1998). Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive. In E.M. Voorhees & D.K. Harman (Eds.), Proceedings of the 7th text retrieval conference (TREC-7), NIST special publication 500–242 (pp. 253–264), Gaithersburg, MD: NIST.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of information by computer. Reading, MAS: Addison-Wesley.

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to passage retrieval in full text information systems. In R. Korfhage, E. Rasmussen, & P. Willet (Eds.), Proceedings of the 16th annual international ACM-SIGIR conference on research and development in information retrieval, Pittsburgh, PA (pp. 49–58), New York: ACM.

Salton, G., Allan, J., & Singhal, A.K. (1996). Automatic text decomposition and structuring. Information Processing and Management, 32(2), 127–138.

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

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

Salton, G., Singhal, A.K., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. Information Processing and Management, 33(2), 193–207.

Singhal, A.K. (1997). AT&T at TREC-6. In E.M. Voorhees & D.K. Harman (Eds.), Proceedings of the 6th text retrieval conference (TREC-6), NIST special publication 500–240 (pp. 215–226), Gaithersburg, MD: NIST.

Singhal, A.K., Buckley, C., & Mitra, M. (1996a). Pivoted document length normalization. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), Proceedings of the 19th annual international ACM-SIGIR conference on research and development in information retrieval, Zurich, Switzerland (pp. 21–29), New York: ACM.

Singhal, A.K., Choi, J., Hindle, D., Lewis, D.D., & Pereira, F. (1998). AT&T at TREC-7. In E.M. Voorhees & D.K. Harman (Eds.), Proceed-

ings of the 7th text retrieval conference (TREC-7), NIST Special Publication 500–242 (pp. 239–252), Gaithersburg, MD: NIST.
Singhal, A.K., Salton, G., Mitra, M., & Buckley, C. (1996b). Document length normalization. Information Processing and Management, 32(5), 619–633.
Stanfill, C., & Waltz, D.L. (1992). Statistical methods, artificial intelligence, and information retrieval. In P.S. Jacobs (Ed.), Text-based intelligent systems: Current research and practice in information extraction and retrieval (pp. 215–225). Hillsdale, NJ: Lawrence Erlbaum Associates.
van Rijsbergen, C.J. (1979). Information retrieval. London: Butterworths.
Voorhees, E.M., & Harman, D. (1996). Overview of the fifth text retrieval conference (TREC-5). In D.K. Harman (Ed.), Proceedings of the 5th text retrieval conference (TREC-5), NIST special publication 500–238 (pp. 1–28), Gaithersburg, MD: NIST.
Voorhees, E.M., & Harman, D. (1997). Overview of the sixth text retrieval conference. In E.M. Voorhees & D.K. Harman (Eds.), Proceedings of the 6th text retrieval conference (TREC-6), NIST special publication 500–240 (pp. 1–24), Gaithersburg, MD: NIST.
Voorhees, E.M., & Harman, D. (1988). Overview of the seventh text retrieval conference (TREC-7). In E.M. Voorhees & D.K. Harman (Eds.), Proceedings of the 7th text retrieval conference (TREC-7), NIST special publication 500–242 (pp. 1–24), Gaithersburg, MD: NIST.
Walker, S., Robertson, S.E., Boughanem, M., Jones, G.J.F., & Sparck-Jones, K. (1997). Okapi at TREC-6 Automatic ad hoc, VLC, routing, filtering and QSDR. In E.M. Voorhees & D.K. Harman (Eds.), Proceedings of the 6th text retrieval conference (TREC-6), NIST special publication 500–240 (pp. 125–136), Gaithersburg, MD: NIST.
Wilkinson, R. (1994). Effective retrieval of structured documents. In B.W. Croft & C.J. van Rijsbergen (Eds.), Proceedings of the 17th annual international ACM-SIGIR conference on research and development in information retrieval, Dublin, Ireland (pp. 311–317), New York: ACM.
Xu, J., & Croft, B.W. (1996). Query expansion using local and global document analysis. In H.-P. Frei, D. Harman, P. Schäuble, & R. Wilkinson (Eds.), Proceedings of the 19th annual international ACM-SIGIR conference on research and development in information retrieval, Zurich, Switzerland (pp. 4–11), New York: ACM.
Yaari, Y. (1997). Segmentation of expository texts by hierarchical agglomerative clustering. In Proceedings of the conference on recent advances in natural language processing Tzigov Chark, Bulgaria (pp. 59–65).
Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments. In B.W. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), Proceedings of the 21st annual international ACM-SIGIR conference on research and development in information retrieval, Melbourne, Australia (pp. 307–314), New York: ACM.
Zobel, J., & Moffat, A. (1998). Exploring the similarity space. SIGIR forum, 32(1), 18–34.
Zobel, J., Moffat, A., Wilkinson, R., & Sacks-Davis, R. (1995). Efficient retrieval of partial documents. Information Processing and Management, 31(3), 361–377.