Comparison of Fragmentation Schemes for Document Retrieval

Ross Wilkinson   Justin Zobel
ross@cs.rmit.oz.au   jz@cs.rmit.oz.au
Department of Computer Science, RMIT
GPO Box 2476V, Melbourne, Victoria, Australia 3001

Abstract

Information systems usually rank whole documents to identify which are answers. However, it may in some circumstances be more appropriate to rank fragments to identify which documents are answers. We consider methods of fragmenting and examine the retrieval effectiveness achieved by each method.

1 Introduction

Document structure can be used to improve retrieval effectiveness.[2,4–6] When retrieving only titles or abstracts, document size is highly constrained, as is the content, so that each document is likely to consider only one topic. Thus the whole of the document is likely to be equally relevant to the query. In contrast, a longer document such as a technical article or news report is likely to address several topics.

However, the usual description of a document as a set of weighted terms can lead to the fine structure of a document being completely lost, since the presence a highly relevant fragment will be obscured by other, irrelevant material. It is therefore preferable to match the fragments themselves. This leads to two key problems: what should the fragments be, and how should relevance of fragments be used to determine the relevance of whole documents.

In this paper we consider the first of these questions, and examine several different strategies for dividing documents into fragments. Our conclusion is that the best retrieval is achieved when fragments are constructed to be of similar length, and that there may be a small advantage to fragments that are part of a logical unit such as a section. For the purposes of this discussion, we take the view that if one fragment of the document is of interest, the whole document should be identified as an answer. That is, we have not considered how information about relevance of fragments could be combined to give relevance information about documents.

2 Document fragmentation

There have been several proposed methods for forming fragments. Obvious choices are sentences, paragraphs and pages. The minimum and maximum size of these fragments is another factor that can be explored. Salton and colleagues[1,5] have demonstrated that the use of individual sentences can help determine the relevance of whole documents. In last year's TREC experiments, we demonstrated that pages of 1,000 bytes are useful fragments for retrieval.[8] In order to ensure that poor breaking of a document did not occur, Callan[2] demonstrated that overlapping of fragments can be successful. We have also investigated the explicit structure that can be provided in documents by SGML markup. Having fragmented using this structure, we investigated how the similarity of these fragments could be combined with the structural information to determine the relevance of whole documents,[6,8] with mixed results.

Innovative strategies have been pursued by other researchers. Hearst and Plaunt[3] have shown how structure can be discovered and has advocated queries based on content in context. Mittendorf and Schauble[4] showed how hidden Markov models can be used to discover passages that can be used as retrieval fragments.

We compare several approaches: retrieval based on whole documents; retrieval based on sentences; retrieval based on paragraphs, of a range of fixed maximum lengths; retrieval based on pages, that is, sequences of paragraphs of similar length; and retrieval based on sections, of a range of fixed minimum and maximum lengths.

3 Experimental design

The database used for these experiments was the set of 45,820 documents that comprise the Federal Register (FR) subset of the TREC collection. This set of documents is a good test-bed as it shows a wide variation in size, and documents have an explicit structure, marked up in SGML. The document collection has 70 million words in total, and so average document length is about 1,500 words.

In all experiments, the documents were stemmed but not stopped, and were stored and retrieved using the MG text retrieval system.[7] This system implements a cosine similarity measure and is returns a ranked list with associated similarities. The similarity measure is given by

$$C(q,d) = \frac{\sum_{t \in q \wedge d}(w_{q,t} \cdot w_{d,t})}{\sqrt{\left(\sum_{t \in q}(w_{q,t})^2 \cdot \sum_{t \in d}(w_{d,t})^2\right)}}$$

with

$$w_{x,t} = \log(f_{x,t} + 1) \cdot (\log(N/f_t) + 1)$$

where $f_{x,t}$ is the frequency of t in x, N is the number of documents in the collection, and f_t is the number of documents containing t.

In the case of retrieving documents using fragments, the frequency information is based on the number of fragments containing each term, and a document's similarity is determined by the similarity of the highest ranked fragment.

The query set for these experiments is the subset of 54 of the TREC topics 51-150 that have relevant FR documents. These topics have the SGML mark-up removed and are stopped and stemmed. The standard TREC relevance judgements have been used. After processing, each query had on average 90 words.

When dividing documents into sentences, we only considered sentences within the TEXT or LP fields. A sentence break was registered by any of ".", ", ", ".\n", "\n\n", "?", or "!". In the case of a ".", there was a check to ensure that the previous character is not uppercase, to stop initials or acronyms from causing sentence breaks. Using this definition, there were 3,780,437 sentences.

To break documents up into paragraphs, an SGML tag or "\n\n" was taken to represent the end of a paragraph. Pages are defined to be sequences of paragraphs such that each sequence exceeds, but is is as close as possible to, the required page size. Using a page size of 1,000 bytes, there were 367,141 pages.

Documents were split into sections based on the documents' internal markup. The documents contained several tags that defined internal structure. We found that that only the T2 and T3 tags could be reliably used to indicate a new internal fragment. Section breaks were defined to be a blank line, or a line containing only markup, followed by a T2 or a T3 tag. This led to a database of 340,287 sections.

Our experiments investigate the interplay between size and type of fragment. We have previously determined that 1,000 bytes produces a good size for page based retrieval. Maximum paragraph sizes of 100, 500, 1,000, and 5,000 bytes were investigated. Short paragraphs were formed by splitting. We also conducted a similar investigation of section size, testing the effect of imposing minimum sizes and maximum sizes separately, and then combined. Minimum sizes ranged from 100 to 5,000 bytes and maximum sizes from 500 to 10,000 bytes.

4 Results

Our results are shown in Tables 1 to 4. In each table we have shown average precision for the 54 queries at various numbers of documents retrieved; thus for example the first column of numbers in each table is precision at 5 documents. Note that if a document had several fragments with high similarity, then only the highest ranked one was considered, and in general more than 200 fragments must be ranked in order to find the 200 highest ranked documents.

Table 1 describes the effectiveness of retrieving paragraphs of varying maximum size compared with the baseline of retrieving whole documents. We also compare division into pages and sentences.

Table 2 describes the effectiveness of retrieving sections, and then imposing a minimum number of bytes on the sections. Table 3 shows the results of applying a maximum. Thus in the first case we agglomerate sections, in the second case we split sections.

Table 4 shows the best of the methods to date, as well as describing the consequences of putting both minimum and maximum constraints on the size of sections.

It is worth reiterating that these experiments are designed to investigate appropriate fragmentation methods for retrieval, not to determine an appropriate retrieval formula. There has been no attempt to combine evidence, use structural

	Precision at number of documents						
Experiment	5	10	15	25	30	50	200
Documents	0.244	0.228	0.196	0.170	0.147	0.120	0.056
Sentences	0.128	0.092	0.075	0.071	0.063	0.054	0.030
Para-max-100	0.170	0.126	0.114	0.097	0.078	0.060	0.030
Para-max-500	0.093	0.080	0.070	0.064	0.056	0.047	0.026
Para-max-1,000	0.075	0.064	0.059	0.053	0.053	0.043	0.026
Para-max-5,000	0.077	0.063	0.058	0.055	0.054	0.042	0.025
Pages	0.300	0.235	0.206	0.182	0.152	0.117	0.052

Table 1: Comparison of informal fragmentation methods

	Precision at number of documents						
Experiment	5	10	15	25	30	50	200
Sections	0.137	0.113	0.102	0.096	0.078	0.060	0.037
Section-min-100	0.174	0.144	0.131	0.118	0.094	0.078	0.043
Section-min-500	0.233	0.209	0.183	0.159	0.135	0.111	0.053
Section-min-1,000	0.263	0.217	0.200	0.178	0.148	0.117	0.054
Section-min-5,000	0.252	0.233	0.214	0.181	0.157	0.124	0.055

Table 2: Comparison of section fragmentation with minimum size

information, filter documents against minimum similarity criteria.

5 Analysis and conclusions

The first set of experiments show that pages are better units of retrieval than either paragraphs or sentences. Ensuring that paragraphs are not too large makes little difference. The paging method performs significantly better than document ranking in the case of precision at at 5 documents retrieved; but otherwise, although paging has a numerical advantage, these methods cannot be separated by the Wilcoxon signed pairs test.

Ensuring that sections are not too small is clearly beneficial. Even imposing a minimum size of 5,000 bytes—which is quite large—leads to significantly better performance at most levels of retrieval. Curiously, ensuring that sections are no bigger than 1,000 bytes also gives improved retrieval performance; however, this improvement is significant only when retrieving 15 documents, and performance is not as good as with the other methods discussed above. It is interesting to compare having a maximum size for sections and a maximum size for paragraphs: the former gives better retrieval than the latter at almost all numbers of documents retrieved.

A comparison of the best methods shows that there is no significant difference between page-level retrieval and section-level retrieval with a minimum size constraint. Document-level retrieval also works well. The significance of these results is that, for the FR collection, there is only minimal performance gain to be had by fragmenting and using a simple document ranking formula. However, it is interesting to observe that, using fragments as a basis for retrieval, the longer relevant documents are more likely to be identified as answers. That is, use of fragments helps eliminate bias in the cosine method against long documents.

Moreover, in this paper we have concentrated on appropriate fragmenting schemes for ranking. There is clearly potential to better exploit this fine structure. Of particular interest to us is that the section fragmentation now performs as well as any other fragmentation method, but provides the opportunity of using type information, as well as structural information, in order to improve document retrieval.

Acknowledgements

We would like to thank Daniel Lam and Tim Shimmin for their assistance with the experi-

	Precision at number of documents						
Experiment	5	10	15	25	30	50	200
Sections	0.137	0.113	0.102	0.096	0.078	0.060	0.037
Section-max-500	0.159	0.143	0.120	0.110	0.093	0.068	0.033
Section-max-1,000	0.177	0.155	0.141	0.123	0.102	0.075	0.039
Section-max-5,000	0.167	0.139	0.122	0.105	0.084	0.063	0.036
Section-max-10,000	0.167	0.130	0.119	0.105	0.083	0.064	0.036

Table 3: Comparison of section fragmentation with maximum size

	Precision at number of documents						
Experiment	5	10	15	25	30	50	200
Documents	0.244	0.228	0.196	0.170	0.147	0.120	0.056
Pages	0.300	0.235	0.206	0.182	0.152	0.117	0.052
Section-min-1,000	0.263	0.217	0.200	0.178	0.148	0.117	0.054
Section-max-1,000	0.177	0.155	0.141	0.123	0.102	0.075	0.039
Section-1,000-5,000	0.285	0.226	0.200	0.180	0.146	0.117	0.053

Table 4: Comparison of best fragmentation methods

ments described here. We would like to acknowledge the valuable environment of the Multimedia Database Systems group at the Collaborative Information Technology Research Institute. This work was supported by the Australian Research Council and the Centre for Intelligent Decision Systems.

References

[1] J. Allan, C. Buckley, and G. Salton. Automatic routing and ad-hoc retrieval using SMART: TREC 2. In *Proc. Text Retrieval Conference (TREC)*, Gaithersburg, Maryland, 1994. NIST Special Publication 500-215.

[2] P. Callan. Passage-level evidence in document retrieval. In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 302–309, Dublin, Ireland, 1994.

[3] M.A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 59–68, Pittsburg, 1993.

[4] E. Mittendorf and P. Schäuble. Document and passage retrieval based on hidden markov models. In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 318–327, Dublin, Ireland, 1994.

[5] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 49–58, Pittsburg, 1993.

[6] R. Wilkinson. Effective retrieval of structured documents. In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 311–317, Dublin, Ireland, 1994.

[7] I.H. Witten, A. Moffat, and T.C. Bell. *Managing gigabytes: Compressing and indexing documents and images.* Van Nostrand Reinhold, New York, 1994.

[8] J. Zobel, A. Moffat, R. Wilkinson, and R. Sacks-Davis. Efficient retrieval of partial documents. *Information Processing & Management.* (To appear).