Generic Summaries for Indexing in Information Retrieval

Tetsuya Sakai†‡
†Knowledge Media Laboratory
Toshiba Corporate R&D Center, Japan

Karen Sparck Jones‡
‡Computer Laboratory
University of Cambridge, UK

ABSTRACT

This paper examines the use of generic summaries for indexing in information retrieval. Our main observations are that: (1) With or without pseudo-relevance feedback, a summary index may be as effective as the corresponding fulltext index for *precision-oriented search of highly relevant documents*. But a reasonably sophisticated summarizer, using a compression ratio of 10-30%, is desirable for this purpose. (2) In pseudo-relevance feedback, using a summary index at initial search and a fulltext index at final search is possibly effective for *precision-oriented search, regardless of relevance levels*. This strategy is significantly more effective than the one using the summary index only and probably more effective than using summaries as mere term selection filters. For this strategy, the summary quality is probably not a critical factor, and a compression ratio of 5-10% appears best.

1. INTRODUCTION

Although automatic summarization is one way of overcoming information overload problems, evaluating summarization systems can be difficult [8]. An *intrinsic* evaluation raises some issues, since this involves subjective assessment of the summary quality or the assumption of a unique "ideal" summary. An alternative is to conduct an *extrinsic* evaluation, which examines the usefulness of summaries for a particular task, e.g. information retrieval (IR).

In the context of IR, summarization has at least two possibilities: it can be used in *preprocessing* or in *postprocessing*. Many researchers have dealt with the latter, by providing the user with *query-focused* summaries of retrieved documents for efficient relevance preview [1, 5, 12, 22, 24]. In contrast, little work has been reported regarding the former, which involves the use of *generic* summaries instead of fulltexts as the source of *document index terms* to improve retrieval effectiveness and/or efficiency. (For convenience, we treat "full text" as a single word in this paper.) While generic summaries may generally be less effective than query-focused summaries for relevance judgements and relevance feedback, they may still be useful in terms of versatility and reusability. Morever, it seems that the above-mentioned *summary indexing* approach is becoming more important with the rapid growth of the World Wide Web and digital libraries, since previous work suggests that it is particularly suitable for *precision-oriented search* [1].

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA..

To explore the possible uses of summary indexing in IR, this paper tests two hypotheses.

Hypothesis 1 A summary index can be as effective as the corresponding fulltext index for retrieval of highly relevant documents.

This is based on the view that a good generic summary should retain the main contents of the original document while discarding peripheral ones, and that it should be as good a piece of evidence as the original document for determining whether it is highly relevant or not to a certain search request. In this paper, a *highly relevant* document is defined as (1) one that is relevant to a search request according to binary relevance assessment as in TREC [20]; *and* (2) one whose text as a whole, or at least most part of it, is relevant to this request.

Hypothesis 2 A summary index can be useful for pseudo-relevance feedback.

Pseudo-relevance feedback (PRF) [15, 18, 26] *assumes* the initial top ranked documents to be relevant, and therefore its success depends heavily on the initial search precision at high ranks. If summary indexing is suitable for precision-oriented search as mentioned earlier, replacing the fulltext index with a summary index at initial search may boost the number of truly relevant documents in the pseudo-relevant document set. Furthermore, for each initial search term and each candidate expansion term, the number of pseudo-relevant summaries that contain a term (and therefore the number of those which do not) would be different from the fulltext case. This would affect *query expansion* and *term reweighting* [10, 16], and may have a positive effect on PRF.

Since a summary can either be an *extract*, typically obtained through sentence selection, or a more sophisticated *abstract* obtained through some text manipulation [8], we prepared a test collection that has a manual abstract and several automatic extracts for each fulltext. Specifically, we generated extracts of three types, each with five different *compression ratios* (See Section 2.3). By comparing summaries with fulltexts, extracts with abstracts, and extracts of different types and lengths, we could examine the impact of summary information and quality on retrieval performance. In total, our summary indexing experiments involve 66 runs using 17 indexes.

The remainder of this paper is organized as follows: Section 2 distinguishes this paper from previous work. Section 3 describes our summary indexing experiments. Section 4 provides the results and discusses their implications. Finally, Section 5 summarizes our findings.

2. COMPARISONS WITH PREVIOUS WORK

2.1 Summary Indexing

Since Luhn suggested the use of summaries for indexing in IR in 1958 [6], the only substantial study on summary indexing to date appears to have been that by Brandow *et al.* in 1995 [1]. Their results, namely, that summary indexing improves precision at the cost of a large loss in recall, were based on experiments with 12 simple Boolean queries and some 20,000 generic news articles, with "lead" and *tf-idf*-type extracts with 60, 150 and 250 words. In a smaller study, applying the vector-space model for ranking documents, Sumita and Iida [21] used 10 simple phrasal requests and 600 newspaper articles from the BMIR-J1 Japanese test collection [13]. Their summaries included "lead" and *tf-idf*-type extracts containing up to 3 sentences. Based on this small-scale experiment, they claimed that summary indexing is effective for retrieving (usually highly relevant) *A-relevant* documents *and* for filtering out *B-relevant* documents. Mochizuki *et al.* [7] used the same data in a similar experiment, but without utilizing the relevance levels. Their extracts included "lead", *tf-idf*-type and lexical chain extracts, at a compression ratio of 30%. All of their summary indexes hurt precision at all recall levels.

Our experiments were more extensive than these: (1) We used a larger test collection; (2) We considered the use of PRF; (3) We tested extracts at five different compression ratios; (4) We included manual abstracts. However, as we want human-readable generic summaries, we excluded keywords/phrases-style summaries [7, 21]. We used the probabilistic retrieval model [10, 11], but our findings probably apply to other ranking systems as well.

2.2 Summaries for Feedback

Local context analysis (LCA) [26] can be regarded as a kind of PRF that uses the top ranked *passages* instead of top ranked *documents* to provide expansion terms. These passages are text windows of fixed size, and can be regarded as crude query-focused summaries. Strzalkowski *et al.* [23] present the user with query-focused passage extracts from top ranked documents, to augment the initial natural language request (or *topic* [20]) before query development.

Both approaches use *query-focused* summaries as the source of *expansion terms* (i.e. as *filters* for excluding terms extracted from nonrelevant passages). In contrast, our experiments use *generic* summaries as the primary source of *index terms*, together with conventional PRF to test **Hypothesis 2**. Using a summary index at initial search in PRF should: (1) yield a set of pseudo-relevant documents that is different from the fulltext case (*document refinement effect*); and (2) affect query expansion and term reweighting (*term refinement effect*). Since our methods select new terms from summary texts and thus subsume the filter effect, we have done additional *summary filter* runs to separate the index and filter effects. Thus our experimental runs can be categorized as follows:

Single search runs use either a *summary* index or the *fulltext* index only, with no PRF.

Summary-summary/fulltext-fulltext runs use either a *summary* index or the *fulltext* index only, for initial search, PRF, and final search.

Summary-fulltext runs use a *summary* index for initial search and PRF, but the *fulltext* index for final search.

Summary filter runs use the *fulltext* index for initial search, PRF and final search, but *summaries* are used as term selection filters.

The document refinement effect could be separeted from the term refinement effect by using a summary index just to obtain a pseudo-relevant set and then the fulltext index for PRF, but this is beyond the scope of our experiments. Nevetheless, our experiments suggest that the term refinement effect is in fact the dominant factor (See Section 4.1).

2.3 Compression Ratio

The *compression ratio* (CR) of a summary is the ratio of the summary length to the corresponding fulltext length, as measured in characters [22], words [1] or sentences [5]. Morris *et al.* reported that their 20% and 30% summaries were as informative as the fulltexts for a task of solving English reading comprehension questions [9]. Although this work is sometimes misquoted as if they discovered the *optimal* CR for any kind of text for any purpose [1, 27], it is not clear that Shannon's 25% "redundancy" limit from information theory [19] applies to text summarization, or whether reading comprehension questions cover all informative content. Moreover, since the CR can affect summary evaluation [5], we examine extracts at 5%, 10%, 30% and 50% CR. We also generate *abstract-length* extracts, which have the same number of *sentences* as the corresponding manual abstracts, so as to separate the effect of abstraction/extraction from that of varying CR.

3. EXPERIMENTS

3.1 Test Collection

We formed our own test collection using documents, search requests (*topics*) and relevance assessments from TREC test collections as follows.

1. We selected all documents that contain *both* the SUMMARY field and the TEXT field from the set of 293,121 Ziff-Davis technical articles on computing contained in TREC CDs 1 through 3 [20]. Since the SUMMARY field contains a manual Abstract (hereafter simply referred to as "abstract") and the TEXT field contains a Fulltext, this set of 38,884 documents is referred to as the ZIFF-AF document collection [14, 16].

2. We selected all requests to which at least 5 ZIFF-AF documents are relevant from the ad hoc requests of TREC1 through 3 (i.e. from *topics* 51 through 200). There were 30 such requests [14, 16].

Table 1: Document statistics.

fulltext (f)	#char	#word	#sent
length mean	6862.4	1114.6	63.9
manual abstract (a)	#char	#word	#sent
length mean	771.2	118.8	5.9
CR (%) mean	18.1	**17.3**	16.1
CR (%) max	197.6	201.3	250.0
CR (%) min	0.05	0.08	0.04
CR (%) std dev.	9.9	9.7	10.1

Table 2: TITLE+DESCRIPTION lengths.

	mean	max	min
(1) #words in request	21.9	47	8
(2) #terms in query	8.7	20	2

Table 3: Relevant documents per request.

	mean	max	min
#all relevant docs	61.5	260	5
#highly relevant docs	30.2	131	3
%highly relevant docs	56.1	100	17.9

Table 4: Indexes based on automatic extracts.

index	CR (%) mean	#word mean	#sent mean
lead			
ld(5)	3.5	44.5	3.3
ld(10)	8.1	99.0	6.2
ld(AL)	**15.7**	110.5	5.9
ld(30)	28.2	322.3	18.3
ld(50)	48.2	545.6	30.7
tfidf			
tf(5)	5.8	54.8	1.7
tf(10)	8.4	98.7	3.3
tf(AL)	**24.1**	171.6	5.9
tf(30)	28.0	321.1	12.5
tf(50)	48.3	545.7	23.4
1p_tfidf			
1p(5)	2.9	40.2	2.1
1p(10)	7.6	95.0	4.0
1p(AL)	**21.3**	150.8	5.9
1p(30)	27.9	320.8	13.2
1p(50)	48.2	545.5	23.9

3. For each of the above requests, the first author selected highly relevant documents from the set of TREC relevant documents. There were 25 requests for which at least 3 documents were highly relevant. These are referred to as the ZIFF-AF test requests.

Table 1 provides the statistics of the ZIFF-AF document collection. Hereafter, we use *word-based* CR since it seems fairly representative. As the "CR max/min" rows suggest, the collection contains 9 documents whose abstract CRs are over 100% and 91 documents whose abstract CRs are below 1%. These are due to huge lists or tables (e.g. telephone directories) in the *original* fulltext: some were discarded for their inclusion in the TREC data (causing CR max's), while others were retained (causing CR min's). However, these extremely rare cases are highly unlikely to affect our IR experiments.

Table 2 Row (1) provides the length statistics of the ZIFF-AF test requests. Although its request set is rather small for IR evaluation [2], ZIFF-AF test collection is the only one available to us with multiple relevance levels *and* with both an abstract and a fulltext for each document.

Table 3 shows the number of all/highly relevant documents and the percentage of highly relevant documents per request. The highly relevant documents were selected only by the first author, but we have assumed this was adequate for our purpose because: (1) there is evidence that "comparative evaluation of retrieval performance is stable despite substantial differences in relevance judgments" [25]; (2) selecting highly relevant documents from the given TREC relevance sets is probably much easier, and less controversial, than initial relevance assessment. (We found a few *clearly* nonrelevant documents in the TREC relevance sets, which we naturally did not include our highly relevant sets.)

3.2 Indexes

The fulltext and the abstract indexes (hereafter denoted by f and a, respectively) were respectively generated from the TEXT and SUMMARY fields for each ZIFF-AF document. The *extract* indexes, with five different CRs, were generated for each of the following three *extract types*:

lead Sentences were taken one by one from the beginning of the TEXT field [1].

tfidf After stopword removal and stemming, sentences were extracted in decreasing order of the *sentence scores*, which are calculated as the sum of *tf-idf* weights as in [27].

1p_tfidf Sentences were extracted as in **lead** until the whole first paragraph had been extracted. Then, additional sentences were extracted as in **tfidf** from the remaining paragraphs.

For each extract, sentences were extracted so that the *actual* CR does not exceed the *desired* CR, but with a minimum of one sentence. (All indexes also use the TITLE fields, which have been ignored when calculating CR.)

Table 4 provides the length statistics of our extract indexes. Here, for example, the extract indexes based on **lead**, **tfidf**, and **1p_tfidf** with a desired CR of 5% are denoted by ld(5), tf(5) and ld(5), respectively. The indexes labelled with "(AL)" were created using the *abstract-lengh* (AL) extracts. Table 4 shows that the mean CRs of tf(AL) and 1p(AL) are greater than that of the a index (See Table 1) and that of ld(AL), which reflects the fact that **tfidf** favours long sentences. Comparing the vocabulary sets of the a and the AL indexes reveals that the intersection between an abstract with an AL extract is around 40% on average.

3.3 Retrieval using Okapi

All of our experiments used the *Okapi Basic Search System*, based on the *BM25* term weighting formula with Okapi defaults [11]. The initial *queries* were generated from the TITLE and the DESCRIPTION fields of the TREC requests by removing strings such as "Document will discuss" and then applying the normal Okapi stopword list and stemming. Table 2 Row (2) provides the length statistics of our queries. The details of our PRF algorithm, which involves both query expansion and term reweighting, can be found in [10, 16]. The PRF parameters, namely, the number of pseudo-Relevant documents (R), the number of Terms added to the initial query (T) and the downweighting factor for the expansion terms (w), were optimized using the f index, and were fixed to $R = 5, T = 30, w = 0.1$ for *all* runs. Thus, our evaluation of summary-based PRF runs may be viewed as "pessimistic" since the PRF parameters are not necessarily optimal for the summary cases.

The *fulltext-fulltext* (i.e. traditional PRF) run will be denoted by f+f. Other PRF runs will be denoted in a similar manner: for example, ld+ld(5) is a *summary-summary* run that uses the ld(5) index, while ld+f(5) is the corresponding *summary-fulltext* run. Whereas, a *summary filter* run using the ld(5) extracts as filters will be denoted by f/ld+f(5), to reflect the fact that it uses the f index throughout.

4. RESULTS AND DISCUSSIONS

Section 4.1 examines single search runs and summary-summary runs and mainly discusses **Hypothesis 1**. Section 4.2 examines the summary-fulltext runs and summary filter runs, and discusses **Hypothesis 2**. In both sections, we use TREC Average Precision (AveP) based on the top 1,000 document ranks and Precision at Document cutoff 10 (PDoc10), the latter for precision-oriented search. These are computed for all/highly relevant documents. (We did full sets of significance tests based on these measures using the Sign Test, but there is no room for the tables here: they appear in our Technical Report [17]. In reporting specific comparisons here, we will therefore refer to the relevant table in [17].) Finally, Section 4.3 compares our selected runs using *discounted cumulative gain* (DCG) curves which make explicit use of different relevance levels for IR evaluation [4].

4.1 Single Search and Summary-Summary Runs

Tables 5–8 show the results of our single search runs (i.e. runs *without* PRF) and summary-summary runs (i.e. runs *with* PRF). Each column represents a CR: in particular, the "AL" (abstract-length) column covers abstracts and AL extracts, while the "100%" column represents fulltexts, equivalent to extracts at 100% CR. In Tables 5–8,

(1) the **boldface** values imply that they are at least as effective on average as the *fulltext baseline*, namely, f (i.e. fulltext performance without PRF) for the single search runs, and f+f (i.e. fulltext performance with PRF) for the summary-summary runs;

(2) for each extract type, the highest performance value among all CRs (except for AL, which is somewhat artificial) is marked with a "$\star$";

(3) the *summary-summary* values are marked with a "$\Downarrow$" if they are below the corresponding *single search* values, that is, if PRF hurts performance on average.

For example, Table 7 provides the following information regarding tf+tf(30) for PDoc10 with all relevant documents: (1) it is not as effective as f+f; (2) it outperforms tf+tf(5), tf+tf(10) and tf+tf(50); (3) it is not as effective as tf(30).

4.1.1 *Fulltexts vs Abstracts vs Extracts*

The lack of **boldface** values in Tables 5 and 6 suggests that, with AveP, summaries are not as effective as fulltexts, regardless of summary types (i.e. extracts or abstracts), of different CRs, of whether PRF is applied or not, and of relevance levels. In particular, Table 6 appears to challenge **Hypothesis 1**, which claimed that summaries are as effective as fulltexts for retrieving highly relevant documents. Moreover, some of the differences between fulltexts and extracts (but not abstracts) are statistically significant for AveP with highly relevant documents (See Table 13(1) and (2) in [17]).

In contrast, the **boldface** values in Table 8 suggest that **Hypothesis 1** may possibly be true for PDoc10, i.e. that a summary index may be as effective as a fulltex index for *precision-oriented search* of highly relevant documents. It seems that 10-30% extracts may be as effective as fulltexts and that abstracts may in fact be better for this particular task, both for the single search and the summary-summary cases. The differences between fulltexts and abstracts/extracts are *not* statistically significant for PDoc10 with highly relevant documents, which is also in agreement with **Hypothesis 1** (See Table 13(1) and (2) in [17]).

In Tables 5–8, abstracts generally outperform extracts regardless of extract types, of different CRs (including AL extracts), of whether PRF is applied or not, of relevance levels and of performance measures. Moreover, some of the differences between abstracts and 5-10%/AL extracts are statistically significant (See Table 14(1) and (2) in [17]). Thus the summary quality seems to be one important factor for both the single search case and the summary-summary case. If a reasonably sophisticated summarizer is used for precision-oriented search of highly relevant documents, it could probably support retrieval performance somewhere between those of abstracts and extracts shown in Table 8.

Regarding **Hypothesis 2**, the single search results in Table 7 (except that with the a index) seem to challenge one assumption underlying it, namely, that a summary index may yield a better set of pseudo-relevant documents for PRF than a fulltext index does. Thus any gain obtained by the use of an extract index with PRF in our experiments is probably attributable to the *term refinement effect* rather than the *document refinement effect* (See Section 2.2).

Effect of CR

As the "$\star$"'s in Tables 5 and 6 suggest, the AveP values increase monotonically with CR, regardless of extract types, of whether PRF is applied or not, and of relevance levels. In contrast, the "$\star$"'s in Tables 7 and 8 indicate that this is not true for precision-oriented search. Although this may be partly because PDoc10 is a less stable measure than AveP [2], the consistency of this trend across Tables 7 and 8 as well as Tables 11 and 12 (See Section 4.2) suggests that a CR of 30% or below may indeed be best for precision-oriented search using a summary index. The significance test results support this view to some extent: for example, for PDoc10 with all relevant documents, tf(30) significantly outperforms tf(5) and tf(10) while tf(50) does not, and tf+tf(30) significantly outperforms tf+tf(5) while tf+tf(50) does not (See Table 15(1) and (2) in [17]).

Effect of Extract Types

The performance ranking of the three extract types is generally not consistent across CRs and across tables, and most

Table 5: Single search/Summary-summary AveP for all relevant documents.

		strategy	5%	10%	AL	30%	50%	100%
fulltext baseline		f	—	—	—	—	—	0.256
single-search	manual abstracts	a	—	—	0.186	—	—	—
	automatic extracts	ld	0.101	0.116	0.132	0.172⋆	0.172⋆	—
		tf	0.121	0.138	0.169	0.196	0.219⋆	—
		1p	0.105	0.130	0.165	0.197	0.219⋆	—
fulltext baseline		f+f	—	—	—	—	—	0.278
summary-summary	manual abstracts	a+a	—	—	0.193	—	—	—
	automatic extracts	ld+ld	0.099⇓	0.131	0.137	0.184	0.205⋆	—
		tf+tf	0.118⇓	0.143	0.178	0.201	0.219⋆	—
		1p+1p	0.111	0.134	0.169	0.205	0.227⋆	—

Table 6: Single search/Summary-summary AveP for highly relevant documents.

		strategy	5%	10%	AL	30%	50%	100%
fulltext baseline		f	—	—	—	—	—	0.249
single-search	manual abstracts	a	—	—	0.231	—	—	—
	automatic extracts	ld	0.147	0.159	0.186	0.215⋆	0.215⋆	—
		tf	0.155	0.166	0.198	0.216	0.236⋆	—
		1p	0.144	0.162	0.192	0.219	0.233⋆	—
fulltext baseline		f+f	—	—	—	—	—	0.276
summary-summary	manual abstracts	a+a	—	—	0.235	—	—	—
	automatic extracts	ld+ld	0.130⇓	0.174	0.196	0.224	0.238⋆	—
		tf+tf	0.149⇓	0.172	0.212	0.218	0.241⋆	—
		1p+1p	0.148	0.163	0.193	0.223	0.238⋆	—

Table 7: Single search/Summary-summary PDoc10 for all relevant documents.

		strategy	5%	10%	AL	30%	50%	100%
fulltext baseline		f	—	—	—	—	—	0.360
single-search	manual abstracts	a	—	—	**0.380**	—	—	—
	automatic extracts	ld	0.276	0.296	0.308	0.320⋆	0.320⋆	—
		tf	0.280	0.280	0.332	0.344⋆	0.320	—
		1p	0.260	0.280	0.300	0.348⋆	0.328	—
fulltext baseline		f+f	—	—	—	—	—	0.360
summary-summary	manual abstracts	a+a	—	—	**0.396**	—	—	—
	automatic extracts	ld+ld	0.268⇓	0.308	0.332	0.340	0.356⋆	—
		tf+tf	0.280	0.324	**0.368**	0.340⋆ ⇓	0.336	—
		1p+1p	0.300	0.304	0.356	0.348⋆	0.332	—

Table 8: Single Search/Summary-summary PDoc10 for highly relevant documents.

		strategy	5%	10%	AL	30%	50%	100%
fulltext baseline		f	—	—	—	—	—	0.252
single-search	manual abstracts	a	—	—	**0.304**	—	—	—
	automatic extracts	ld	0.240	**0.252**⋆	**0.260**	0.248	0.248	—
		tf	0.228	0.236	0.236	**0.252**⋆	0.224	—
		1p	0.220	0.236	0.224	**0.256**⋆	0.236	—
fulltext baseline		f+f	—	—	—	—	—	0.260
summary-summary	manual abstracts	a+a	—	—	**0.316**	—	—	—
	automatic extracts	ld+ld	0.240	**0.260**	**0.280**	0.256	**0.284**⋆	—
		tf+tf	0.232	**0.264**⋆	**0.276**	0.248⇓	0.240	—
		1p+1p	0.252	0.248	**0.284**	**0.264**⋆	0.232⇓	—

of the differences across extract types are *not* statistically significant (See Table 16(1) and (2) in [17]).

Thus the findings in Section 4.1 can be summarized as follows: *With or without pseudo-relevance feedback (i.e. single search or summary-summary), a summary index may be as effective as the corresponding fulltext index for precision-oriented search of highly relevant documents. But a reasonably sophisticated summarizer, using a compression ratio of 10-30%, is desirable.*

4.2 Summary-Fulltext and Summary Filter Runs

Tables 9–12 show the results of our summary-fulltext runs (i.e. PRF runs that use a summary index except for final search) and summary filter runs (i.e. PRF runs that use summaries only as filters). The conventions for **boldface**, "$\star$" and "$\Downarrow$" are as before. In addition, in Tables 9–12,

- the values are marked with a "$\downarrow$" if they are below the corresponding *summary-summary* values;
- the *summary-fulltext* values are marked with a "$\otimes$" if they are below the corresponding *summary filter* ones.

For example, in Table 12, the lack of "$\otimes$"'s implies that all summary-fulltext runs are at least as effective on average as the corresponding summary filter runs, while "$\Downarrow\downarrow$" in the "AL" column indicates that f/a+f is less effective on average than a and a+a, for PDoc10 with highly relevant documents.

4.2.1 Fulltexts vs Abstracts vs Extracts

Tables 9 and 10 suggest that a+f, f/a+f and even some of the extract-based runs are comparable to the fulltext baseline for AveP. With PDoc10, the results appear to be even more promising despite our "pessimistic" evaluation (See Section 3.3). In particular, the **boldface** values in Tables 11 and 12 indicate that all summary-fulltext runs outperform f+f regardless of relevance levels (except tf+f(50)), suggesting that the summary-fulltext strategy is effective for precision-oriented search. However, as these precision gains are *not* statistically significant (See Table 13(3) in [17]), our results do not fully confirm **Hypothesis 2**, namely, that a summary index is useful for PRF. Using the summary-fulltext strategy for precision-oriented search nevertheless deserves further investigation with test collections with larger search request sets.

On the other hand, Tables 11 and 12 also suggest that the summary quality is *not* a critical factor for the summary-fulltext strategy to achieve high precision, since, for example, 1p+f(5) is as effective as a+f in these tables. Indeed, for PDoc10, there is no significant difference between abstracts and extracts for the summary-fulltext case, regardless of relevance levels (See Table 14(3) in [17]). Thus, *if* the summary-fulltext strategy is indeed effective for precision-oriented search, probably any existing summarizer would be sufficient for this purpose.

4.2.2 Summary-Fulltext vs Summary Filter

As the "$\otimes$"'s in Tables 9 and 10 and the lack of them in Tables 11 and 12 show, summary filter runs generally outperform the corresponding summary-fulltext runs with AveP, while all summary-fulltext runs outperform the corresponding summary filter runs for PDoc10. As discussed in Section *4.2.1*, the summary-fulltext results for PDoc10 are probably the ones with practical importance, since their gain over the fulltext baseline appears to be substantial.

The differences between summary-fulltext runs and summary filter runs are in general *not* statistically significant (See Table 18(3) in [17]). However, the following significance test results indirectly suggest that summary-fulltext runs are more promising for precision-oriented search than summary filter runs, and therefore that a summary index is more than just a term selection filter:

1. With PDoc10, summary-fulltext runs are more successful with PRF than summary filter runs, regardless of relevance levels (See Table 17(2) and (3) in [17]).
2. With PDoc10, while many summary-fulltext runs significantly outperform the corresponding summary-summary runs for all relevant documents, no summary filter runs do (See Table 18(1) and (2) in [17]).
3. With PDoc10, while the differences between abstracts and extracts for the summary-fulltext case are *not* statistically significant regardless of relevance levels, some of the differences for the summary filter case *are* significant (See Table 14(3) and (4) in [17]). That is, the summary filter runs are less robust with respect to summary quality.

In addition, the differences between summary-fulltext runs and summary-summary runs are statistically significant for AveP as well, especially for all relevant documents (See Table 18(1) and (2) in [17]). The superiority of summary-fulltext runs over summary-summary runs is also indicated by the lack of "$\downarrow$"'s alongside the summary-fulltext values in Tables 9–12.

In short, the summary-fulltext strategy is significantly more effective than the summary-summary strategy regardless of performance measures, and probably more effective than the summary-filter strategy for precision-oriented search.

Effect of CR

As the "$\star$"'s in Tables 9 and 10 suggest, the "monotonic increase" property for AveP generally holds for summary filter runs, while this is not true for summary-fulltext runs. On the other hand, for PDoc10, Tables 11 and 12 suggest that 5-10% CR is best for summary-fulltext runs, and that 30% CR is best for summary filter runs. Significance tests suggest that summary filter runs are more sensitive to CR than summary-fulltext runs (See Table 15(3) and (4) in [17]).

Effect of Extract Types

Most of the differences due to extract types are *not* statistically significant (See Table 16(3) and (4) in [17]). Thus we cannot conclude as to which extract type is the most effective from our experiments.

Thus the findings in Section 4.2 can be summarized as follows: *In pseudo-relevance feedback, using a summary index at initial search and a fulltext index at final search (i.e.*

Table 9: Summary-fulltext/Summary filter AveP for all relevant documents.

		strategy	5%	10%	AL	30%	50%	100%
fulltext baseline		f+f	—	—	—	—	—	0.278
summary-fulltext	manual abstracts	a+f	—	—	**0.284**	—	—	—
	automatic extracts	ld+f	0.250⊗	0.259⊗	0.259⊗	0.255⊗	0.262⋆⊗	—
		tf+f	0.267	**0.280**⋆	0.268⊗	0.270⊗	0.256⊗	—
		1p+f	0.271	0.260⊗	0.272⊗	0.273⋆⊗	0.264⊗	—
summary filter	manual abstracts	f/a+f	—	—	**0.279**	—	—	—
	automatic extracts	f/ld+f	0.268	0.274	0.274	0.275	**0.278**⋆	—
		f/tf+f	0.267	0.270	0.274	0.277	**0.278**⋆	—
		f/1p+f	0.262	0.272	0.274	0.277	**0.278**⋆	—

Table 10: Summary-fulltext/Summary filter AveP for highly relevant documents.

		strategy	5%	10%	AL	30%	50%	100%
fulltext baseline		f+f	—	—	—	—	—	0.276
summary-fulltext	manual abstracts	a+f	—	—	**0.287**	—	—	—
	automatic extracts	ld+f	0.238⊗	0.259⋆⊗	0.266⊗	0.256⊗	0.257⊗	—
		tf+f	0.253⊗	0.270⋆	0.252⊗	0.266⊗	0.266⊗	—
		1p+f	0.251⊗	0.238⊗	0.265⊗	0.271⋆⊗	0.262⊗	—
summary filter	manual abstracts	f/a+f	—	—	**0.278**	—	—	—
	automatic extracts	f/ld+f	0.264	0.271	0.271	0.273	**0.277**⋆	—
		f/tf+f	0.260	0.259	0.271	0.274	**0.276**⋆	—
		f/1p+f	0.256	0.261	0.271	0.275	**0.276**⋆	—

Table 11: Summary-fulltext/Summary filter PDoc10 for all relevant documents.

		strategy	5%	10%	AL	30%	50%	100%
fulltext baseline		f+f	—	—	—	—	—	0.360
summary-fulltext	manual abstracts	a+f	—	—	**0.436**	—	—	—
	automatic extracts	ld+f	**0.380**	**0.420**⋆	**0.388**	**0.400**	**0.396**	—
		tf+f	**0.416**	**0.424**⋆	**0.416**	**0.384**	**0.368**	—
		1p+f	**0.436**⋆	**0.404**	**0.420**	**0.404**	**0.372**	—
summary filter	manual abstracts	f/a+f	—	—	**0.380**↓	—	—	—
	automatic extracts	f/ld+f	0.348	**0.360**	**0.380**	**0.376**⋆	**0.364**	—
		f/tf+f	0.340	0.340	0.356↓	**0.368**⋆	**0.360**	—
		f/1p+f	0.340	0.336	**0.360**	**0.368**⋆	**0.360**	—

Table 12: Summary-fulltext/Summary filter PDoc10 for highly relevant documents.

		strategy	5%	10%	AL	30%	50%	100%
fulltext baseline		f+f	—	—	—	—	—	0.260
summary-fulltext	manual abstracts	a+f	—	—	**0.324**	—	—	—
	automatic extracts	ld+f	**0.284**	**0.312**⋆	**0.300**	**0.288**	**0.296**	—
		tf+f	**0.300**⋆	**0.296**	**0.304**	**0.280**	0.256	—
		1p+f	**0.324**⋆	**0.304**	**0.308**	**0.296**	**0.268**	—
summary filter	manual abstracts	f/a+f	—	—	**0.280**⇓↓	—	—	—
	automatic extracts	f/ld+f	0.252	**0.268**⋆	**0.272**↓	**0.268**⋆	**0.264**↓	—
		f/tf+f	0.244	0.244↓	**0.260**↓	**0.264**⋆	0.256	—
		f/1p+f	0.248↓	0.240↓	**0.264**↓	**0.264**⋆	0.256	—

summary-fulltext strategy) is possibly effective for precision-oriented search, regardless of relevance levels. This strategy is significantly more effective than the one using the summary index only (i.e. summary-summary strategy) and probably more effective than using summaries as mere term selection filters (i.e. summary filter strategy). For the summary-fulltext strategy, the summary quality is probably not a critical factor, and a compression ratio of around 5-10% appears best.

4.3 Discounted Cumulative Gain Curves

While previous sections used traditional IR performance measures and treated all/highly relevant documents separately, DCG curves give us a unified, intuitive picture. For our DCG-based evaluation of selected runs, we assigned a gain of 2 to each highly relevant document, and 1 to each *partially relevant* (i.e. relevant but not highly relevant) document. Log with base 2 was used for discounting the cumulative gain, to reflect the fact that documents at low ranks are less valuable to the user in general [4]. Figure 1 shows the DCG curves of f, a and tf(30), bearing on **Hypothesis 1**, and f+f, a+f and tf+f(10), bearing on **Hypothesis 2**, as well as part of the theoretical "best possible" curve.

The DCG curves show that, while a is at least as effective as f, tf(30) is not quite so even at document ranks 1–10. This appears contradictory to our finding related to **Hypothesis 1**, that a summary index may be effective for precision-oriented search of *highly* relevant documents. But this is probably because DCG curves treat *all* relevant documents as at least somewhat relevant: the DCG curves of f and tf(30) actually become much closer to each other if we assign a gain of 0 to each partially relevant document or if we assign a gain that is higher than 2 to each highly relevant document. (The practical importance of a highly relevant document may indeed be more than twice that of a partially relevant document in the case of our TREC data.) Nevertheless, as discussed in Section 4.1, the summary quality does seem to be an important factor here. Possibly, the DCG curve of a more sophisticated summarizer may be comparable to that of f, or even of a.

On the other hand, it can be observed that the DCG curves of a+f and tf+f(10) are above that of f+f, which is in agreement with Tables 11 and 12. Thus, both PDoc10 and DCG curves appear to support **Hypothesis 2**, namely, that a summary index can be useful for PRF. However, as mentioned in Section *4.2.1*, more experiments are required in order to show the advantage of the summary-fulltext strategy over traditional PRF in terms of statistical significance.

5. CONCLUSIONS

This paper has examined the use of generic summaries for indexing in information retrieval. Our main observations are that:

1. With or without pseudo-relevance feedback, a summary index may be as effective as the corresponding fulltext index for *precision-oriented search of highly relevant documents*. But a reasonably sophisticated summarizer, using a compression ratio of 10-30%, is desirable for this purpose.

2. In pseudo-relevance feedback, using a summary index at initial search and a fulltext index at final search is possibly effective for *precision-oriented search, regardless of relevance levels*. This strategy is significantly more effective than the one using the summary index only and probably more effective than using summaries as mere term selection filters. For this strategy, the summary quality is probably not a critical factor, and a compression ratio of 5-10% appears best.

In order to obtain more concrete evidence and more general conclusions, our experiments should be repeated using test collections with larger search request sets, with multiple relevance levels, and, if possible, with both an abstract and a fulltext for each document. Above all, the possible advantage of the summary-fulltext strategy over traditional PRF for precision-oriented search should be re-examined: This strategy already has an advantage in terms of *efficiency*, since its initial search and query expansion stages are faster than those of traditional PRF, particularly if the compression ratio is as small as 5-10%. Thus it may be a practical solution for tasks such as searching the World Wide Web.

6. ACKNOWLEDGMENTS

The authors are grateful to Stephen Walker and Stephen Robertson for their useful advice.

7. REFERENCES

[1] Brandow, R., Mitze, K. and Rau, L. F: Automatic Condensation of Electronic Publications by Sentence Selection, *Information Processing and Management* 31(5), pp. 675–685 (1995).

[2] Buckley, C. and Voorhees, E.: Evaluating Evaluation Measure Stability, *ACM SIGIR 2000 Proceedings*, pp. 33–40 (2000).

[3] Goldstein, J., Kantrowitz, M., Mittal, V. and Carbonell, J.: Summarizing Text Documents: Sentence Selection and Evaluation Metrics, *ACM SIGIR '99 Proceegings*, pp. 121–128 (1999).

[4] Järvelin, K. and Kekäläinen, J.: IR Evaluation Methods for Retrieving Highly Relevant Documents, *ACM SIGIR 2000 Proceedings*, pp. 41–48 (2000).

[5] Jing, H., Barzilay, R., McKeown, K. and Elhadad, M.: Summarization Evaluation Methods: Experiments and Analysis, *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pp. 60–68 (1998).

[6] Luhn, H. P.: The Automatic Creation of Literature Abstracts, *IBM Journal of Research & Development* 2(2), pp. 159–165 (1958).

[7] Mochizuki, H., Iwayama, M. and Okumura, M.: Document Retrieval using Excerptions (in Japanese), *Workshop on the Current Status and the Future of Text Summarization at the 4th Annual Meeting of the Association for Natural Language Processing*, pp. 22–29 (1998).

[8] Mani, I. and Maybury, M. T.(eds.): *Advances in Automatic Text Summarization*, MIT Press (1999).

[9] Morris, A. H., Kasper, G. M. and Adams, D. A.: The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance, *Information Systems Research* 3(1), pp. 17–35 (1992).

[10] Robertson, S. E. and Sparck Jones, K: Simple, Proven Approaches to Text Retrieval, TR356,

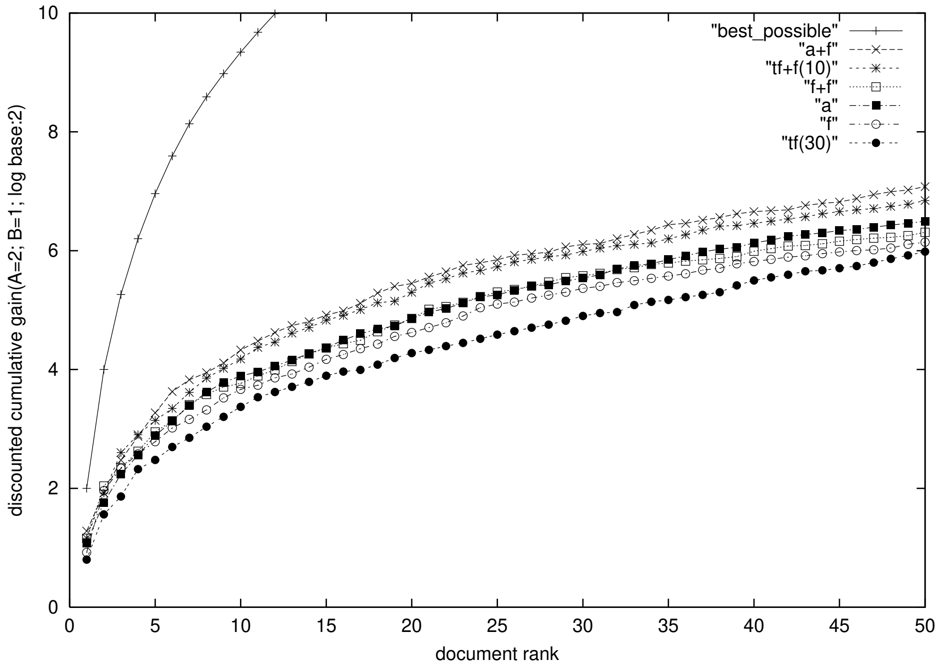

Figure 1: Selected DCG curves.

Computer Laboratory, University of Cambridge (1997). http://www.ftp.cl.cam.ac.uk/ftp/papers/reports/#TR356

[11] Robertson, S. E. and Walker, S.: Microsoft Cambridge at TREC-9, *TREC-9 Proceedings* (2001).

[12] Sakai, T., Miike, S. and Sumita, K.: Evaluation of the Abstract Presentation Interface for a Document Retrieval System (in Japanese), *Information Processing Society of Japan SIG Notes*, HI–57–7, pp. 49–54 (1994).

[13] Sakai, T. *et al.* : BMIR-J2: A Test Collection for Evaluation of Japanese Information Retrieval Systems, *ACM SIGIR Forum*, Vol. 33, No. 1, pp. 13–17 (1999). http://www.acm.org/sigir/forum/F99/tetsuya.sakai.pdf

[14] Sakai, T.: MT-based Japanese-English Cross-Language IR Experiments using the TREC Test Collections, *IRAL 2000 Proceedings*, pp. 181–188 (2000).

[15] Sakai, T., Robertson, S. E. and Walker, S.: Flexible Pseudo-Relevance Feedback via Direct Mapping and Categorization of Search Requests, *BCS-IRSG ECIR 2001 Proceedings*, pp. 3–14 (2001).

[16] Sakai, T.: Japanese-English Cross-Language Information Retrieval using Machine Translation and Pseudo-Relevance Feedback, *International Journal of Computer Processing of Oriental Languages*, Vol. 14, No. 2 (2001).

[17] Sakai, T. and Sparck Jones, K.: Generic Summaries for Indexing in Information Retrieval - Detailed Test Results, TR513, Computer Laboratory, University of Cambridge (2001). http://www.ftp.cl.cam.ac.uk/ftp/papers/reports/#TR513

[18] Sakai, T. and Robertson, S. E.: Flexible Pseudo-Relevance Feedback Using Optimization Tables, *ACM SIGIR 2001 Proceedings* (2001).

[19] Shannon, C. E.: Prediction and Entropy of Printed English, *Bell Systems Technical Journal* 30, pp. 50–64 (1951). pp. 194–208 (1951).

[20] Sparck Jones, K.: Further Reflections on TREC, *Information Processing and Management* 36, pp. 37–85 (2000).

[21] Sumita, E. and Iida, H.: Information Retrieval using a Statistical Abstraction Method (in Japanese), *Proceedings of the 3rd Annual Meeting of the Association for Natural Language Processing*, pp. 353–356 (1997).

[22] SUMMAC: http://www-nlpir.nist.gov/related_projects/tipster_summac/final_rpt.html (1998).

[23] Strzalkowski, T., Wang, J. and Wise, B.: Summarization-based Query Expansion in Information Retrieval, *COLING-ACL '98 Proceedings*, pp. 1258–1264 (1998).

[24] Tombros, A. and Sanderson, M.: Advantages of Query Biased Summaries in Information Retrieval, *ACM SIGIR '98 Proceedings*, pp. 2–10 (1998).

[25] Voorhees, E. M.: Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness, *ACM SIGIR '98 Proceedings*, pp. 315–323 (1998).

[26] Xu, J. and Croft, W. B.: Query Expansion Using Local and Global Document Analysis, *ACM SIGIR '96 Proceedings*, pp. 4–11 (1996).

[27] Zechner, K.: Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences, *COLING '96 Proceedings*, Vol. 2, pp. 986–989 (1996).