TREC 7 Ad Hoc, Speech, and Interactive tracks at MDS/CSIRO

Michael Fuller* Marcin Kaszkiel* Dongki Kim†‡ Corinna Ng* John Robertson†§ Ross Wilkinson†¶
Mingfang Wu* Justin Zobel*

1 Overview

For the 1998 round of TREC, the MDS group, long-term participants at the conference, jointly participated with newcomers CSIRO. Together we completed runs in three tracks: ad-hoc, interactive, and speech.

2 Ad-hoc task

In TREC-5 we used document retrieval based on arbitrary passages [8, 9], or fixed-length passages that could start at any word position. Although far from the best runs in TREC-5, these results were promising, in particular for long documents. In TREC-6 we continued with arbitrary passages, but our main emphasis was on comprehensive factor analysis of successful automatic query expansion and refinements methods in the context of the vector space model [5]. This year we have refined the MG retrieval system to include Rocchio-based relevance feedback. Also, phrase matching has been added. We have continued to use arbitrary passages and combination of evidence for document retrieval.

2.1 System description

An in-house version of the MG retrieval system has been used for all experiments. All experiments were carried out on an Intel Pentium II (300 Mhz) with a single processor and 256 Mb of physical memory.

Queries and documents were matched using the Okapi formulation [13]:

$$sim(q,d) = \sum_{t \in q \wedge d} w_{d,t} \cdot w_{q,t} \quad (1)$$

with $w_{d,t}$:

$$\frac{(k_1+1)\cdot f_{d,t}}{k_1 \cdot [(1-b) + b \cdot \frac{W_d}{avr\_W_d}] + f_{d,t}}$$

and $w_{q,t}$:

$$\frac{(k_3+1)\cdot f_{q,t}}{k_3 + f_{q,t}} \cdot \log \frac{N - f_t + 0.5}{f_t + 0.5}$$

where k_1, k_3, and b are constants set to 1.2, 1000, and 0.75 respectively, as recommended by the City University group [13]. The value W_d is the length of document d in bytes and avr_W_d is the average document length in the entire collection. The value N is the total number of documents in the collection, f_t is the number of documents in which term t occurs, and $f_{x,t}$ is the frequency of term t in either document d or query q.

Okapi is not easily adaptable to arbitrary passage ranking because parameters k_x and b are tuned to document ranking. Queries and passages are matched using a non-normalised version of the cosine similarity function:

$$sim(q,p) = \sum_{t \in q \wedge p} (w_{q,t} \cdot w_{p,t}) \quad (2)$$

with weights that have been shown to be robust and give good retrieval performance [1]: $w_{q,t} = (log(f_{q,t}) + 1) \cdot log(\frac{N}{f_t} + 1)$ and $w_{p,t} = log(f_{p,t}) + 1$.

Automatic relevance feedback was based on the Rocchio formula [12]:

$$Q_{new} = \alpha \cdot Q_{orig} + \frac{\beta}{|R|} \sum_{r \in R} r + \frac{\gamma}{|R'|} \sum_{r' \in R'} r' \quad (3)$$

where Q_{orig} is a weighted term vector for the original query; R is the set of relevant documents; R' is the set of non-relevant documents; and r and r' are weighted term vectors for relevant and non-relevant document, respectively. Parameters α, β, and γ control the contribution of terms from original query, relevant documents, and non-relevant documents, respectively.

For indexing purposes, documents and queries have been stopped using the stop-list used in our TREC-6 experiments.[1] Single terms have been stemmed with the Lovins algorithm [10]. Two-word phrases are indexed if they satisfy the following conditions:

- individual words of the phrase occur at least 30 times in collection, and
- the phrase occurs at least 10 times in collection.

A detailed description of two-term phrase extraction can be found elsewhere [3].

2.2 Ad-hoc runs

This year we have concentrated on short queries, and have submitted official runs for *title* and *title+description* queries. For the first time we have not submitted a full-topic run.

* Department of Computer Science, RMIT,
GPO Box 2476V, Melbourne VIC 3001, Australia
{msf,martin,cln,ross,ming,jz}@mds.rmit.edu.au
† CSIRO, Division of Mathematical and Information Science
‡ GPO Box 664, Canberra ACT 2601, Australia
Dong.Ki.Kim@cmis.csiro.au
§ Locked Bag 17, North Ryde, NSW 1670, Australia
John.Robertson@cmis.csiro.au
¶ 723 Swanston St, Carlton VIC 3053, Australia
Ross.Wilkinson@cmis.csiro.au

[1] See Appendix A of the MDS TREC-6 report for a list of stopped terms [5].

	5 docs	10 docs	20 docs	200 docs	Avg. Prec.	%Δ
title						
Document	0.424	0.382	0.332	0.124	0.161	0.0
Passage-300	0.424	0.388	0.322	0.127	0.162	+0.6
mds98t	**0.436**	**0.422**	**0.365**	**0.159**	**0.220**	+36.6
mds98t2	**0.440**	**0.426**	**0.359**	**0.159**	**0.218**	+35.4
mds98t-p300	0.432	0.404	0.355	0.159	0.218	+35.4
title+desc						
Document	0.532	0.486	0.397	0.145	0.204	0.0
Passage-300	0.556	0.458	0.375	0.140	0.194	-4.9
mds98td	**0.572**	**0.536**	**0.446**	**0.187**	**0.281**	+37.7
mds98td-p300	0.540	0.508	0.423	0.180	0.261	+27.9
title+desc+narr						
Document	0.580	0.536	0.450	0.167	0.240	0.0
Passage-300	0.524	0.472	0.394	0.154	0.214	-10.8
mds98tdn	0.616	0.554	0.483	0.196	0.285	+18.8
mds98tdn-p300	0.580	0.518	0.444	0.190	0.271	+12.9

Table 1: *TREC-7 ad-hoc results.*

A complete set of results for TREC-7 is shown in Table 1. For completeness, full-topic runs are included. Official runs are shown in bold face.

Runs *mds98t* and *mds98td*, which correspond to title and title+description queries, used the following approach. Single terms and phrases were used; weights of phrases were scaled down by 0.3 to compensate for single-term contributions of the terms of the phrase. Documents were ranked using the Okapi formulation (equation 1) and 1000 documents retrieved. The top ten documents retrieved (that is, set R) were assumed to be relevant and the last 250 documents of the 1000 retrieved (that is, set R') were assumed to be non-relevant. Using the parameter values $\alpha = 1.0$, $\beta = 2.0$, and $\gamma = 1.0$, equation 3 was used to select an *additional* 40 single terms and an *additional* 5 phrases. Each new single term had to appear in at least 2 relevant documents in order to be considered. With the original query terms re-weighted and new terms added to the query, a final set of 1000 documents was retrieved using document ranking.

Run *mds98t* also a simplified form of this approach: phrases were not used, since there was not enough evidence in very short queries to justify their use; as short queries are not likely to retrieve many relevant documents in top 10, only the first 5 documents were assumed to be relevant; and 80 new terms were added to original query with no restriction of minimum occurrence in relevant documents.

Run *mds98t2* (title queries) is an experimental run that explores combination of evidence, as we have done in past TRECs [5, 8]. The scores of documents in mds98t2 are based on a weighted sum of the document scores from mds98t and the document scores based on the arbitrary passage of 300 words. The document scores for passage-based ranking were downweighted by 0.3 to take into account poorer retrieval performance with respect to the relevance feedback run of mds98t.

It is interesting to observe that, for short queries, document retrieval based on arbitrary passages is as effective as sophisticated document ranking. However, as the query length increases, query terms' proximity is not as important as occurrence of multiple query terms in the entire document.

A two-stage relevance feedback achieved a significant improvement for different types of queries. For title and title+description queries, the improvement for average precision was at least 36%. We believe that the improvement is not as significant for full topic queries because the parameter tuning had been based on queries of medium length.

	Best	≥ Median
mds98t	2	29
mds98t2	1	28
mds98td	4	42

Table 2: *Number of queries that had highest average precision or had at least median average precision.*

Table 2 compares our official runs with the automatic runs submitted by other TREC-7 participants. The table shows the number of queries for MDS runs that achieved at least a median average precision or had the highest average precision for a topic.

For short queries, Table 2 does not reflect well the relative performance because all automatic runs were used to derive the best and median values. Despite this, at least 30 title queries performed better than the median run. For the title+description run, only four queries fell short of the median. This is a positive result and indicates that simple relevance feedback can be highly effective.

To test the robustness of Rocchio-based relevance feedback, the same approach has been used on last year's data. Table 3 compares our TREC-7 strategy with the results from last year's TREC. A 37% increase has been achieved over last year's results and a 25% improvement over the Okapi measure. This result is consistent with our TREC-7 results.

Run mds98t2 fell short of expectations. There was no improvement from combining mds98t with arbitrary passage ranking.

2.3 Further experiments

In order to further evaluate document retrieval based on arbitrary passages, we have used expanded queries from document-based retrieval and ranked documents based on best passage. The average precision for title+description queries has decreased by 7.1% from mds98td but for title queries there was no differences. See Table 1 for run mds98t-p300 and mds98td-p300 (note: phrases were not used in those runs).

We have explored many parameters of the Rocchio formulation on past TREC data. Our TREC-7 runs have used

Experiment	5 docs	10 docs	20 docs	200 docs	Avg. Prec.	%Δ
Last year's	0.400	0.370	0.260	0.120	0.204	0.0
Document	0.464	0.426	0.347	0.128	0.224	+9.8
Passage-150	0.496	0.424	0.331	0.122	0.242	+18.6
TREC7 method	0.530	0.460	0.400	0.159	0.280	+37.3

Table 3: *TREC-6 experiments (title+description queries).*

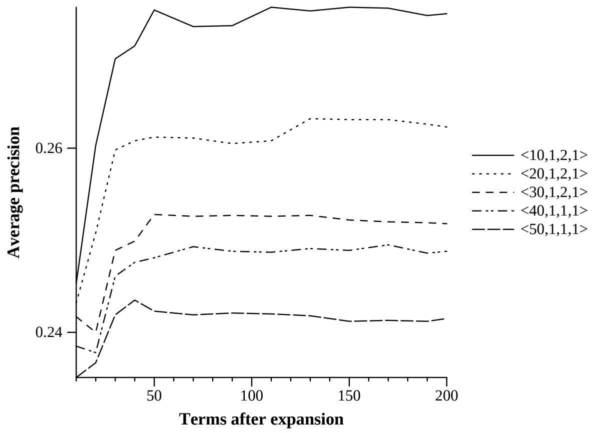

Figure 1: *Rocchio-based relevance feedback on TREC-6 data (title+desc queries).*

parameters that were most consistent over the training data. Here we summarize why particular parameters have been used in TREC-7.

For each R, the best combination of α, β and γ has been identified and presented on single graph. The results are summarised in Figure 1. Each line corresponds to a combination of $\langle |R|, \alpha, \beta, \gamma \rangle$. The figure illustrates that the most consistent results were achieved when R was set to 10, and α, β, and γ set to 1, 2, and 1, respectively. For all cases $|R'|$ was 250; we found in our experiments that negative feedback was useful.

2.4 Analysis

Overall, we conclude the following:

- Using a simple two-stage retrieval, up to 36% gain has been achieved over a single stage retrieval;
- Statistical phrases has improved average precision by 4.1%;
- Document retrieval based on single arbitrary passage is at least as effective as entire document ranking, confirming our past results [9].

3 Interactive retrieval track

The purpose of these experiments was to examine how different organizations of query results affect the ability of users to resolve information needs, focusing on retrieval coverage and efficiency. In particular for TREC-7, the MDS group focused on comparing a *cluster*-based organization with a simple *list*-based organization. Two interactive systems were tested, with the main distinction that where one displayed an ordered list of document titles, the other displayed an ordered list of clusters and their descriptors. The performance of the two systems could be evaluated in two ways: how effective each system was at helping an interactive user identify relevant documents, and how efficient each system was at helping an interactive user identify relevant documents.

To isolate this comparison from other interactive effects, the interfaces of the two systems were kept as consistent as was possible, no querying facility was provided, no relevance feedback was sought, and no query reformulation was possible. As a result, for each topic, the same pool of candidate documents was offered to each subject. Note that these restrictions greatly diminish the role of the user in contrast with other interactive experiments, at the cost of lowering overall performance.

3.1 Retrieval engine

The MG [15] retrieval system was used to identify the pool of candidate documents for each topic. Documents were casefolded, stemmed (Lovins [10]), and stopped.[2] To form queries, the description portion (not title, not narrative) of each TREC-7 topic was case-folded, stemmed, and stopped. Term weights in documents was calculated by using *tf.idf*; term weights in queries used *idf*. Queries were matched against the document collection using the cosine measure. The top 300 ranked documents formed the pool of candidate documents available to the experiment subjects for each topic.

For the clustered-organization, the pool of 300 candidate documents was clustered using two passes of a single-pass clustering algorithm [4, 16]. The number of clusters for each query was controlled between 7 and 10; the size of each cluster was not controlled. Within a cluster, documents were ranked according to their similarity to the query. Cluster descriptors were formed from the ten highest-weighted terms from the cluster vector, the five most frequent word pairs from all documents in the cluster, and the titles of the three documents in the cluster most similar to the query.

3.2 User interfaces

Given the goal of comparing two alternate organizations of the same data, it was important that the two interfaces be as consistent as possible, differing only in their presentation of the alternate organizations. The design of the interfaces also assumed that relatively large monitors would be available for the interactive experiments, sufficient to permit side-by-side viewing of documents and result organizations. For ease of development, a suite of perl CGI scripts was used to generate the HTML and JavaScript that implemented the interfaces. No mechanism for providing relevance feedback or for supplying a new query was provided; subjects were restricted to exploring the pool of pre-selected candidate documents.

For the list-based organization, the viewport was divided into two parts. (See Figure 2.) The left half displayed a

[2] See Appendix A of the MDS TREC-6 report for a list of stopped terms [5].

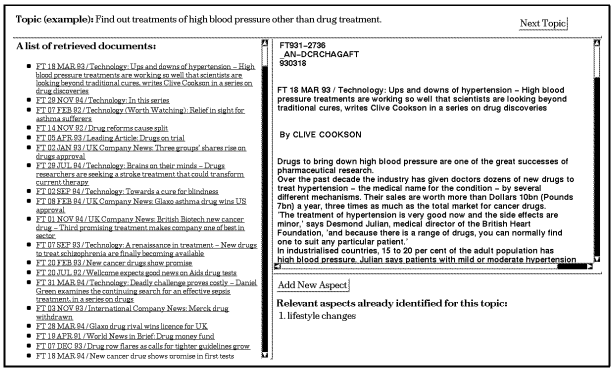

Figure 2: *List-based user interface.*

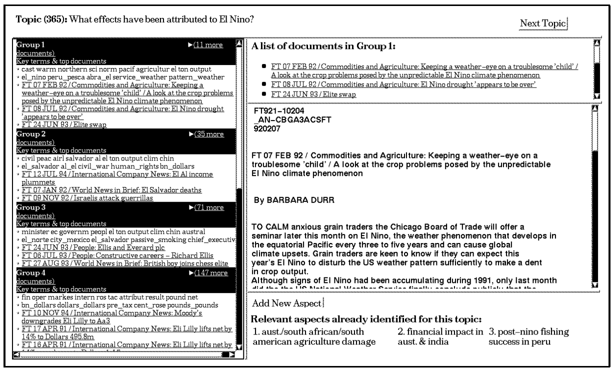

Figure 3: *Cluster-based user interface.*

ranked, scrollable list of the titles of the top 300 documents for a topic; each title could be selected by single-clicking. The upper part of right half of the viewport, initially blank, displayed a scrollable view of any document selected in the left-hand panel. The lower part of right half of the viewport, the aspect selection panel, displayed a list of currently known document aspects; subjects could use this panel to record that a document contained an aspect relevant to the topic and to add a description of the aspect via a pop-up dialogue box.

For the cluster-based organization, the left-hand panel was replaced by an ordered, scrollable list of cluster descriptors; each cluster could be selected by single-clicking. (See Figure 3.) Each cluster descriptor identified the cluster by number, indicated how many documents it contained, provided a list of representative terms and a list of representative term pairs, and listed the three titles of the three 'top' documents from the cluster. The scrollable document view of the list-based interface was divided into two parts. The upper part, initially blank, was used to display a scrollable, ranked list of the titles of documents in any cluster selected in the left-hand panel; each title could be selected by single-clicking. The lower part, initially blank, displayed a scrollable view of any document selected in the upper panel. The aspect selection panel was unchanged.

3.3 Subjects

Sixteen subjects undertook the experiment, according to the Latin Square arrangement stipulated by the TREC-7 Interactive Track guidelines.[3] Their task was "to save documents, which, taken together, contain as many different instances as possible of the type of information the topic expresses a need for—within a 15 minute time limit" from a pool of candidate documents selected from the TREC-7 "Financial Times of London 1991-1994" collection, for each of eight (slightly modified) TREC-7 adhoc topics.

The subjects were undergraduate computer science students, recruited via an internal RMIT newsgroup. The subjects aged from 17 to 23, and had on average 3.3 years of on-line search experience.

The subjects attempted four searches using each system. Pre- and post-experiment, post-search, and post-system questionnaires were administered to each subject, as was the psychometric test FA-1 (Controlled Associations) from ETS "Kit of Reference Tests for Cognitive Factors"' (1976 Edition). Documents identified as relevant by subjects, as well as other "significant" events in each session, were logged and timestamped automatically.

3.4 Results

The effectiveness of the two organizations was measured in terms of aspectual recall and aspect coverage; the efficiency was measured by the time taken to locate each new aspect. The distribution of the assessed aspects for each topic in the pool of candidate documents is shown in Table 4 and Table 5.

To measure effectiveness, the list of documents whose full text was viewed during each search was extracted from the experiment logs. The aspectual recall of this list can be calculated using either the judgement of the independent NIST assessors, as shown in Table 6, or the judgement of the experimental subjects, as shown in Table 7. The assessors' judgement provides an objective assessment of the quality of the documents that were chosen for viewing, whereas the subjects' judgement reflects their own concept of document's relevance to the topic in terms of their own understanding of the information need.

To measure efficiency, the time to locate each aspect was calculated, again according to both the assessors' judgement, as shown in Figure 4, and the subjects' judgement, as shown in and Figure 5.

There were no significant differences for overall aspectual recall between the the two organizations, nor for the time to locate each aspects (paired, one tail t-test).

The pre-experiment psychometric test attempted to gauge subjects' associational fluency in terms of the number of synonyms for a set of eight stimulus words. The mean score for the 16 subjects was 23.2 correct of 34.9 total terms, with a standard deviation of 8.1 terms. There appeared to be a linear correspondence between subjects' FA-1 score and their average aspectual recall; no correlations were found with performance with either interface.

Previous work has indicated that clustering can be an effective mechanism for identifying rich groups of candidate documents [6], and in particular, that clustering can be used to improve ad-hoc query performance in terms of recall-precision [16]. To see if a similar effect would be observed in terms of aspectual recall, we have selected only the documents that are part of clusters from which subjects saved documents. Table 8 shows the aspectual recall of those documents when ordered by their original cluster-ranking and within-cluster ranking. Compared with the baseline (Table 4), aspectual

[3] `http://www.nist.gov/itl/div893/894.02/projects/t7i/`

Topic	Number of documents								
	5	10	15	20	50	100	150	200	300
352	0.000	0.000	0.000	0.036	0.107	0.107	0.107	0.143	0.143
353	0.091	0.182	0.364	0.364	0.364	0.545	0.545	0.545	0.545
357	0.385	0.385	0.462	0.462	0.462	0.615	0.692	0.692	0.692
362	0.000	0.167	0.167	0.167	0.167	0.250	0.333	0.333	0.333
365	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.792	0.792
366	0.000	0.000	0.000	0.000	0.286	0.571	0.571	0.571	0.857
387	0.111	0.111	0.111	0.222	0.556	0.889	0.889	1.000	1.000
392	0.111	0.417	0.472	0.500	0.639	0.750	0.750	0.806	0.806
Avg.	0.181	0.251	0.291	0.312	0.416	0.560	0.580	0.610	0.646

Table 4: *Aspectual recall for the candidate document pool of the 300 highest-ranked documents.*

	Topic number							
	352	353	357	362	365	366	387	392
Total aspects in all documents	28	11	13	12	24	7	9	36
Total documents containing aspects	120	69	86	116	40	39	88	88
Aspects in candidate documents	4	6	9	4	19	6	9	29
Candidate documents containing aspects	3	13	27	10	3	9	35	30

Table 5: *Aspect coverage for each topic, as judged by the NIST assessors.*

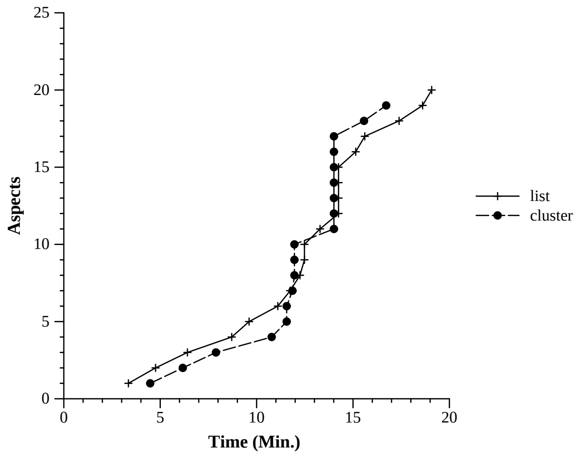

Figure 4: *Time to get each assessor-determined aspect for documents whose full text was viewed.*

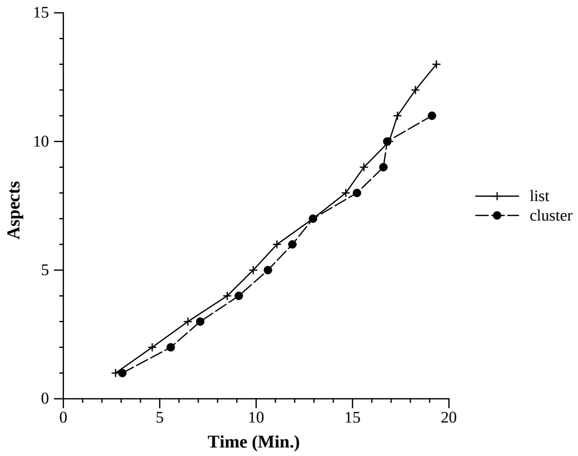

Figure 5: *Time to get each subject-determined aspect for documents whose full text was viewed.*

Topic	Number of docs retrieved 5	10	15	20	50
352	0.058	0.103	0.112	0.121	0.125
353	0.057	0.159	0.205	0.227	0.250
357	0.241	0.270	0.318	0.337	0.346
362	0.073	0.146	0.156	0.156	0.167
365	0.490	0.495	0.495	0.495	0.495
366	0.125	0.304	0.304	0.339	0.411
387	0.028	0.069	0.111	0.153	0.194
392	0.080	0.167	0.215	0.233	0.278
Avg.	0.144	0.214	0.239	0.258	0.283

(a) Cluster organization

Topic	Number of docs retrieved 5	10	15	20	50
352	0.009	0.049	0.058	0.058	0.058
353	0.057	0.114	0.193	0.239	0.250
357	0.318	0.337	0.366	0.366	0.366
362	0.042	0.136	0.146	0.146	0.156
365	0.693	0.693	0.693	0.693	0.693
366	0.071	0.107	0.161	0.197	0.232
387	0.097	0.139	0.236	0.236	0.305
392	0.226	0.382	0.399	0.413	0.420
Avg.	0.189	0.245	0.281	0.293	0.310

(b) List organization

Table 6: *Cumulative aspectual recall (as judged relevant by assessors) for each topic, at increasing numbers of documents viewed, of documents whose full text was displayed. Note that subjects viewed varying numbers of documents for each topic.*

Topic	Number of docs retrieved 5	10	15	20	50
352	0.440	0.688	0.771	0.844	0.875
353	0.320	0.568	0.627	0.669	0.697
357	0.480	0.621	0.701	0.763	0.763
362	0.257	0.632	0.679	0.679	0.798
365	0.369	0.431	0.431	0.431	0.431
366	0.376	0.494	0.607	0.607	0.732
387	0.242	0.575	0.833	0.858	0.929
392	0.355	0.495	0.646	0.705	0.779
Avg.	0.355	0.563	0.662	0.695	0.750

(a) Cluster organization

Topic	Number of docs retrieved 5	10	15	20	50
352	0.179	0.408	0.621	0.662	0.737
353	0.531	0.698	0.760	0.760	0.823
357	0.560	0.742	0.861	0.861	0.874
362	0.354	0.651	0.666	0.682	0.724
365	0.692	0.692	0.692	0.692	0.692
366	0.146	0.271	0.396	0.458	0.521
387	0.285	0.600	0.758	0.783	0.840
392	0.294	0.494	0.578	0.634	0.658
Avg.	0.380	0.570	0.667	0.692	0.734

(b) List organization

Table 7: *Cumulative aspectual recall (as judged relevant by subjects) for each topic, at increasing numbers of documents viewed, of documents whose full text was displayed. Note that subjects viewed varying numbers of documents for each topic.*

recall improves over the first 20 documents; the results are truncated at 20 documents as not all clusters contained 50 documents (although results at higher levels still indicate improved performance).

Although no significant difference in terms of either efficiency or effectiveness was found between organizations for the overall results, inspection of the topic-by-topic results suggested that there was in fact a variation in performance for a subset of the topics. Table 9 shows that, for the four topics for which users of the list organization saved the fewest aspects, users of the cluster organization saved highly significantly more topics (paired, one tail t-test). Conversely, for the four topics for which users of the list organization saved the most aspects, significantly fewer topics were saved by users of the cluster organization. The same result can be observed in the assessor-based aspectual recall levels of the two subgroups of topics, although not at a statistically significant level (see Table 10). Note that the break-up of topics approximately corresponds to average familiarity with each topic, as determined by post-search questionnaire. No equivalent effect was discernible in per-topic efficiency results.

Additionally, although for both list and cluster organizations subjects saved similar numbers of aspects on average (μ of 33.4 for the list, 31.3 for the cluster), subjects' behaviour when using the cluster interface was far more consistent than with the list interface (σ of 19.1 for the list, 8.2 for the cluster). Apparently, when using the cluster interface, the subjects saved on average the same number of aspects for each query, regardless of query familiarity or the number of aspects to be found.

Searchers' comments

From the exit questionnaire, 12 of the 16 subjects preferred the clustered organization to the list organization, and 13 subjects rated the clustered organization as easy to use.

A fairly clear preference for the cluster organization was shown by the subjects, who made comments such as:

> Clustering interface is easier, because it's grouped.
>
> Everything was nicely organized into group and I could skip some stuff and get directly to the point.
>
> It showed me all the list of the topic in a screen.

In contrast, comments on the simple list organization included:

> Too many links in one list.
>
> Everything was just in a list and it was difficult to concentrate on the actual topic.
>
> Long list, sometimes frustrated in couldn't find suitable topic.
>
> Hard to search, depends on the topic.
>
> Very simple interface - no confusion.

However, subjects did note some inadequacies of the cluster organization as implemented, such as:

> The keywords in each group are not clear. They will make users confused for the first time.

—the cluster descriptor terms were stemmed, rather than complete words—and

	Number of docs retrieved			
Topic	5	10	15	20
352	0.036	0.107	0.107	0.143
353	0.091	0.182	0.364	0.364
357	0.308	0.308	0.385	0.385
362	0.167	0.167	0.167	0.167
365	0.750	0.792	0.792	0.792
366	0.286	0.286	0.571	0.714
387	0.111	0.111	0.444	0.444
392	0.111	0.417	0.472	0.472
Avg.	0.233	0.296	0.413	0.435

Table 8: *Aspectual recall (as judged relevant by assessors) for documents from clusters from which a document was saved, at increasing numbers of documents viewed.*

	352	353	365	366	Avg.
List organization	23	21	22	12	19.5
Cluster organization	30	26	32	21	27.3

(a) "Hard" Topics

	357	362	387	392	Avg.
List organization	43	49	27	70	47.3
Cluster organization	35	33	25	48	35.3

(b) "Easy" Topics

Table 9: *Number of aspects saved, per topic.*

The group is not exactly you want.

—presumably indicating either the failure of the clustering algorithm, or the shortcomings of the cluster ranking algorithm.

3.5 Analysis

Although most subjects liked the clustered organization, we did not find any significant difference overall in either effectiveness or efficiency between the ranked list organization and the clustered organization, However, a statistically significant difference was observed in the number of aspects saved when the set of topics was divided into harder and easier groups; this result carried over to aspectual recall for the two groups, albeit not at a statistically significant level. This suggests that the cluster organization may be helpful for "hard" topics, but of less value for "easy" topics; confirming the general validity of this result and characterizing applicable situations must be the subject of further work.

The cluster hypothesis—that relevant documents tend to cluster—has been verified here; however, the algorithm used was not able to cluster documents into topic aspects. Subjects tended to browse all clusters, but generally saved documents from clusters that contained many aspects (as determined by the NIST assessors); in contrast, subjects generally did not save documents from clusters that contained few topic aspects. This suggests that, while clustering helped subjects identify useful groups of documents, without further aid subjects experienced some difficulty in identifying relevant documents. This is borne out by an improvement over the baseline in aspectual recall when considering documents from relevant (suggested by a subject having saved a document) clusters only. Secondary clustering passes, perhaps in the style of Scatter/Gather [6], once users have located clusters of interesting documents may aid the identification of specific relevant documents or distinct aspects.

	352	353	365	366	Avg.
List organization	.053	.068	.667	.071	.215
Cluster organization	.089	.102	.687	.214	.273

(a) "Hard" Topics

	357	362	387	392	Avg.
List organization	.279	.135	.250	.285	.237
Cluster organization	.231	.146	.111	.184	.168

(b) "Easy" Topics

Table 10: *Aspectual recall per topic of saved documents, as judged relevant by assessors.*

4 Spoken document retrieval

For TREC7 we again chose to explore phoneme-based methods for spoken document retrieval (SDR). We believe that the phonetic approach is required for SDR with data sets containing more than one dialect of English. In our case, we are interested in approaches that can manage American, Australian, and British variations of English. MDS, in collaboration with CSIRO, participated in the full SDR run, which included two speech runs, *mds-s1* and *mds-s2*. The first speech run was submitted by the CSIRO team while the other retrieval runs were submitted by the RMIT team.

The two key processes involved in SDR are speech recognition followed by textual retrieval. The recognition process used the HTK toolkit [18]. The documents were recognised as phoneme sequences. For the reference and baseline retrieval runs, the word-based documents and queries were translated to phonemes using the CMU pronunciation dictionary [2]. The document collection contained 100 hours of News Broadcast obtained from LDC. It contained 2866 documents with an average length of 275 words. A set of 23 queries was used for evaluation. The average length of a query was about 16 words.

4.1 Speech recognition system

Based on our decision to use phone models as the basic units for recognition as well as text retrieval, 39 phones in the CMU dictionary were used for training the continuous density, left-to-right HMMs. As manually segmented and labelled American-accented speech data was not available for training those models, five files were arbitrarily selected from each different news program on the TREC-6 spoken document collection and were then partitioned into smaller files. Some of the partitioned files included noise, music and non-speech. These were filtered out. Finally we obtained trainable speech data of 18.4 hours and converted the corresponding word sequence of each speech file into sequences of phones, using the CMU dictionary and their grapheme-to-phoneme software, which were then used for training the models.

The acoustic parameters used were 12 mel-frequency cepstral coefficients, 12 delta coefficients and two normalised log-energy values and were extracted every 10 ms using a window frame of 25 ms (Hamming windows with pre-emphasis). With these acoustic feature vectors, initially 39 context-independent phone models with one mixture were trained and then 1521

mds-r1	Retrieval using reference transcriptions
mds-b1	Retrieval using baseline 1 transcriptions (35% wer)
mds-b2	Retrieval using baseline 2 transcriptions (50% wer)
mds-s2	Retrieval using phoneme-based transcriptions from our own team's phoneme-based recogniser.

Table 11: *Submitted runs for speech retrieval.*

words	Find reports of fatal air crashes
phonemes	F y N D R I P X R T S h V F x T h L E R K R a s I Z
quad-grams	FyND yNDR ... TShV ShVF ... ThLE ... ERKR ... asIZ
quad-grams, bounded	FyND RIPX ... XRTS hV FxTh xThL ER KRas ... asIZ

Table 12: *Example of differences between an unbounded quad-gram and a bounded quad-gram query.*

Total Number of files used for testing	:	960
% Correct	:	53.43 %
% Accuracy	:	43.34 %

Table 13: *Results of phone recognition.*

	Quad-gram queries
mds-r1	0.3107
mds-b1	0.2753
mds-b2	0.1937
mds-s2	0.1063

Table 14: *Average precision of submitted runs.*

right-context dependent models were trained incrementally with the number of mixtures from one to three, depending on the amount of training data. We chose the right-context models because our informal experiments showed that these models produced slightly better results than the left-context models, and triphone models were not trained due to insufficient training data. The right-context models that had less than 100 training speech tokens were cloned from the context-independent models.

For language modelling, the backoff bigram for the right-context dependent phones was computed from the label files used for training described above. Table 13 shows the results of phone recognition, which was over part of the training data consisting of 11,520 files.

Text retrieval was performed on the speech database of TREC-7 and details are in the subsequent sections.

4.2 Spoken document retrieval experiments from RMIT

Four runs were submitted by RMIT, shown in Table 11. Our phoneme-recognised documents, which had an error rate of about 50%, can be said to be highly corrupted with respect to the other types of transcriptions. Tri-grams and quad-grams of the phonetic transcriptions of the documents were formed and combined prior to indexing by our retrieval system. Queries were also translated to phonetic quad-grams such that no quad-grams were created across word boundaries.

Previous experiments using the TREC-6 spoken document collection, as well as subword unit experiments by Ng et al. [11], indicated that both tri-grams and quad-grams performed well on their own. The term weights of the combined tri-gram and quad-gram phonetic transcriptions would result in a retrieved rank set of documents which would be different to that obtained if they were not combined or by combination of indices [7]. Bounded quad-gram queries, where quad-grams were not created cross word boundaries, caused the formation of quad-grams as well as other shorter n-grams. An example of the differences between a bounded quad-gram and an unbounded one is shown in Table 12.

MG [15], developed at RMIT and University of Melbourne, was the retrieval engine used. The version used included the Okapi similarity formulation (equation 1). The occurrence of negative weights if $f_t \geq \frac{N}{2}$ is handled such that f_t is set to $\frac{N}{2} - 1$.

Four runs based on the different versions of the document collections were submitted. Three of the transcriptions, reference, baseline 1, and baseline 2, were word-based, while our speech recognised run was phoneme-based. The word-based documents were translated to phoneme sequences using the CMU pronunciation dictionary [2]. The word-based queries were translated as well. The original pronunciation dictionary contained approximately 118,000 words. For the training collection, about 1350 words were added and a further 2580 words were added to translate the test collection into phoneme sequences. Prior to translation, the documents were neither stopped nor stemmed.

For the submitted runs, tri-grams and quad-grams of the phonetic transcriptions were created. These were combined for each document prior to indexing. The Okapi similarity function described in the previous section was used for retrieval. Table 14 showed the average precision values for the submitted runs. Using the same retrieval system, the average precision values were also obtained for word-based documents and queries as well as stopped queries. Stopped queries were created using a stopped list of 368 words. The average length of the original queries was about 16 words while the average length of a stopped query was down to about 9 words. The results are shown in Table 15. Further experiments showed that optimal average precision values for the combination of tri-gram and quad-gram phonetic transcriptions can be obtained using a combination of the tri-gram and quad-gram query set. The average precision of these is shown in Table 16.

We translated all word-based documents to phoneme sequences and combined the tri-grams and quad-grams of these sequences for each document prior to retrieval. The queries used were the bounded quad-grams of the translated query where they were not created across word boundaries.

The results indicate that phoneme-based retrieval using n-grams is feasible. Compared to retrieval using word-based

	Original Queries	Stopped Queries
Reference	0.4464	0.4542
Baseline 1	0.4049	0.4171
Baseline 2	0.3403	0.3259

Table 15: *Average precision of word-based baseline retrieval experiments.*

	Original Queries	Bounded Queries
Reference	0.3411	0.3290
Baseline 1	0.3071	0.2886
Baseline 2	0.2314	0.2046
Speech	0.0818	0.1050

Table 16: *Average precision of combined tri-gram and quad-gram queries on combined tri-gram and quad-gram documents.*

documents, as shown in Table 15, phonetic n-gram retrieval did not perform well. This was because there was significant loss of contextual information due to the lost of boundary information after the word documents were translated to phoneme sequences. The use of fixed size combinations of tri-gram and quad-gram for the documents and queries also increased the noise of the collection. There was no significant improvement in retrieval performance using stopped queries in the word-based case, and, although results were not shown here, the same was observed for stopped queries translated to n-grams and used for retrieval.

Using a combination of tri-gram and quad-gram, we found that bounded queries did not improve retrieval performance of the phonetic transcriptions of word-based documents, but there was a slight improvement for phoneme-based documents. This meant that documents which were recognised using a word-based recogniser performed better using the phoneme queries where n-grams were created across boundaries; and documents recognised using a phoneme-based recogniser performed better using bounded queries. A typical result comparing bounded and unbounded queries is shown in Table 16.

The combination of tri-grams and quad-grams showed a slight improvement compared to its performance individually, as shown in Table 17. This was due to increases in numbers of relevant documents found using quad-grams although there was also an increase in noisy matches due to tri-grams.

4.3 Spoken document retrieval experiments from CSIRO

CSIRO submitted a single run, *mds-s1*. The retrieval system developed for this experiment was composed of Perl scripts and C programs developed by CSIRO. The mds-s1 run used the recognition stream developed for mds-s2 but applied a different retrieval technique.

Regardless of what speech recognition system or approach is utilised during the recognition phase, errors will occur, resulting in the creation of a recognition stream with an incorrect representation of the voice data. In the mds-sr1 run, we tested the use of approximate string matching to compensate for these index errors and thus improve retrieval performance. Our hypothesis is that, by supplementing establish speech recognition techniques with approximate string matching techniques, improved retrieval performance will result. As discussed above, the phoneme-based recognizer produced significant levels of errors in the recognition stream (53% accuracy). Because of these inaccuracies an exact string matching approach could not identify any occurrences of a sample set of search terms in the spoken document set's recognition stream. During the HMM (HMM) training phase, HTK produces a table containing performance information, known as the confusion matrix. This table contains detailed information about the types of recognition errors produced by the recognizer (phone substitution, insertion or deletion). By utilizing each phone's performance profile, the approximate string-matching task can eliminate unlikely matches, thus improving matching performance and accuracy.

After the recognition process had been completed, the retrieval phase started with a series of perl scripts used to modify and convert both the query and recognition stream's formats. Upon completion of the conversion process, the retrieval system combined the use of approximate string matching with the confusion matrix to identify occurrences of each search term within the recognition stream.

Several steps were used to convert the queries supplied by TREC into a format acceptable to the mds-s1 retrieval system. Initially, SGML tags and punctuation were removed from the query sentences. Non-noun words were removed from the query using the Moby Part-of-Speech Dictionary [14]. The Carnegie Mellon Pronouncing Dictionary [2] was then used to convert the remaining query terms into their phonetic equivalence. The phonetic label set used by the Carnegie Mellon Pronouncing Dictionary varied from the label set used by the mds-s1 retrieval system. For this reason a conversion script was applied to the query set to translate from the Carnegie Mellon based phonetic query terms to a representation appropriate for the mds-s1 matching algorithm. We referred to the query terms produced through this process as a "processed query term".

The file produced by the recogniser contains a continuous stream of labels representing the phonetic content of the voice track. A sliding window was used to move over the continuous phonetic string one label at a time to parse and generate a set of fixed length sub-strings. The size of the sliding window used was the number of characters in the processed query term. Agrep [17] was then employed as a fast filtering tool to quickly eliminate all sub-strings within the phonetic stream that could not contain the phonetic representation of the search term. Since agrep does not provide insight into which edit operations are utilized during its approximate string matching operation, it was necessary to develop a module that would include in its output the number, order, and type of edit operations required to establish a match. The output of this module includes information about which symbols were deleted, which symbols were inserted and which symbols were replaced by another symbol.

The matching algorithm can derive multiple matches for a single recognition sub-string because the same match can be produced using different sets of operations. Therefore, the output of the matching algorithm typically contains numerous occurrences of the same match, with different combinations of insertion, deletion and substitution operations. It was proposed that the greater the match redundancy for a recognition stream sub-string, the more likely the match is correct. An algorithm was devised that takes into account a match's redundancy, a threshold factor, and the number of nouns within the query that were matched for the spoken document, yielding a ranking for each query and each spoken document.

The experimental results are disappointing. Overall average precision for the mds-s1 run was 0.0223. Due to limited time and resources, the reference and baseline runs were not

	Tri-gram		Quad-gram	
	Original Queries	Bounded Queries	Original Queries	Bounded Queries
Reference	0.3065	0.3033	0.3230	0.3005
Baseline 1	0.2599	0.2757	0.2936	0.2531
Baseline 2	0.2051	0.2015	0.2265	0.1871
Speech	0.0421	0.0561	0.1013	0.0978

Table 17: *Average precision of n-gram transcripts using n-gram queries.*

conducted. Preliminary analysis of the results shows that, while recall over the total document set was reasonable (of the 390 relevant documents 268 were retrieved – 68.7%), precision was woefully inadequate. It was recognized at the submission time that the ranking algorithm was deficient.

Acknowledgements

The work reported in this paper has been partially funded by the Cooperative Research Centres Program through the Department of the Prime Minister and Cabinet of Australia, and by the Australian Research Council.

References

[1] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In W. Croft and C. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 292–300, July 1994.

[2] Cmudict.0.4: Carnegie Mellon University Pronouncing Dictionary. Available from: `http://www.speech.cs.cmu.edu/cgi-bin/cmudict/`, 1995.

[3] J. Fagan. *Experiments in Automatic Phrase Indexing for Docuement Retrievel: A comparison of syntactic and nonsyntactic methods.* PhD thesis, Cornel University, September 1987.

[4] W. B. Frakes and R. Baeza-Yates, editors. *Information Retrieval: Data Structures and Algorithms.* Prentice-Hall, Inc., Englewood Cliffs, NJ, U.S.A., 1992.

[5] M. Fuller, M. Kaszkiel, C. L. Ng, P. Vines, R. Wilkinson, and J. Zobel. MDS TREC6 report. In E. M. Vorhees and D. K. Harman, editors, *Proceedings of the Sixth Text Retrieval Conference*, Gaithersburg, MD, U.S.A., 1997. Department of Commerce, National Institute of Standards and Technology.

[6] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 76–84, Zurich, Switzerland, 18-22 August 1996. ACM.

[7] G. J. F. Jones, J. T. Foote, K. S. Jones, and S. J. Young. Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 30 – 38, 1996.

[8] M. Kaszkiel, P. Vines, R. Wilkinson, and J. Zobel. The MDS Experiments for TREC5. In D. Harman, editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 209–216, November 1996.

[9] M. Kaszkiel and J. Zobel. Passage retrieval revisited. In N. J. Belkin, D. Narasimhalu, and P. Willett, editors, *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185, July 1997.

[10] J. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computation*, 11(1-2):22–31, 1968.

[11] K. Ng and V. W. Zue. Subword unit representations for spoken document retrieval. In *Proc. ESCA Eurospeech Conference*, pages 1607 – 1610, Rhodes, Greece, 1997.

[12] G. Salton, editor. *The Smart retrieval system: experiments in automatic document processing.* Prentice-Hall, 1971.

[13] S. Walker, S. Robertson, M. Boughanem, G. Jones, and K. S. Jones. Okapi at TREC-6 Automatic ad hoc, VLC, routing, filtering and QSDR. In E. Voorhees and D. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 125–136, November 1997.

[14] G. Ward. Moby part-of-speech dictionary. `http://www.dcs.shef.ac.uk/research/ilash/Moby/`, 1996.

[15] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and indexing documents and images.* Van Nostrand Reinhold, 1994.

[16] M. Wu and R. Wilkinson. Evaluation of indexing methods for clustering. In J. Kay and M. Milosavljevic, editors, *Proceedings of the Third Australian Document Computing Symposium*, pages 14–19, August 1998. Proceedings available as Technical Report 518, Basser Department of Computer Science, University of Sydney.

[17] S. Wu and U. Manber. Fast text searching with errors. Technical report, Department of Computer Science, University of Arizona, 1991. Technical Report TR-91-11.

[18] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK Book.* Entropic Cambridge Research Laboratory, 1995.