Effectiveness of Clustering in Ad-Hoc Retrieval

David A. Evans, Alison Huettner, Xiang Tong, Peter Jansen, Jeffrey Bennett

CLARITECH Corporation
A Justsystem Group Company

Abstract In this paper, we describe the experiment underlying the CLARITECH entries in the TREC-7 Ad Hoc Retrieval Track. Based on past results, we have come to regard accurate, selective relevance feedback as the dominant factor in effective retrieval. We hypothesized that a clustered rather than a ranked presentation of documents would facilitate judgments of document relevance, allowing a user to judge more documents accurately in a given period of time. This in turn should yield better feedback performance and ultimately better retrieval results. We found that users were indeed able to find more relevant documents in the same time period when results were clustered rather than ranked. Retrieval results from the cluster run were better than results from the ranked run, and those from a combined run were better still. The difference between the ranked and combined runs was statistically significant for both recall and average precision.

1 Introduction

The most successful approaches to ad-hoc retrieval in recent TREC evaluations have typically involved a combination of manual query formulation, interaction with a user to determine some number of candidate relevant documents, and "relevance" feedback to the system for use in expanding a query and automatically generating a final set of ranked documents. Based on our results in TREC 6 in particular [1], we have come to regard accurate, selective relevance feedback as the dominant factor in determining a successful outcome. In any practical system, such relevance feedback depends on the ability of a user to review and judge a sample of documents in a relatively short amount of time.

Virtually all TREC systems that have utilized user feedback have presented the user with "relevance ranked" lists of documents to review. But such lists may not represent unbiased samples of potentially relevant documents. Serial-order presentation may not be the most appropriate way to organize results. Making the user read or browse documents in isolation may not contribute to the user's efficiency, in particular, in deciding whether to continue reviewing documents, to stop, or to reformulate the query and try again.

Our ad-hoc retrieval experiments in TREC 7 were designed to assess the effectiveness of clustered groups of documents as an alternative to relevance-ranked lists in assisting the user in making relevance judgments. The fifty queries (351–400) were entered into the CLARIT system and edited by a member of the team; these constituted a fixed set of initial, manually prepared queries in all subsequent steps. Eight subjects were enlisted as "users". Their task was to submit the initial queries to a database consisting of the target corpus (excluding the Federal Register collection), and judge results. Results were presented either as relevance-ranked lists of 300 returned documents (the baseline or "ranked" run) or as clustered groups of the top 150 returned documents (the "cluster" run). Each user was assigned some number of queries; half were processed as baseline and half as cluster runs. Users' judgments (documents marked relevant, non-relevant, or merely viewed) were automatically collected at 10, 15, 20, and 30 minutes. During the first 15 minutes, no user interactions with the system were allowed except for the reading and marking of documents. Between 15 and 30 minutes, users were also allowed to reformulate queries and retrieve potentially new results. All fifty queries were processed in each mode; each user processed a query only once in one or the other mode.

In terms of efficiency alone, we observed a positive effect for cluster representations. At all collection points (10, 15, 20, and 30 minutes) the average number of positive judgments per query is higher for the cluster mode. For example, the average number of marked-relevant documents at each point is 8.7, 11.8, 13.9, and 18.7 for the baseline and 9.1, 12.6, 15.9, and 20.5 for the cluster runs. In terms of overall performance—average precision and recall—our official results further demonstrate the higher performance of the cluster runs.

In the following sections, we report on our experimental design, the results we obtained in the several modes of processing, our overall performance on the TREC-7 task, and the results of several follow-up analyses we conducted.

2 Experiment Design

The CLARIT TREC-7 ad hoc retrieval experiment was designed to measure the effect of document clustering on the speed and quality of user relevance judgments. To conduct our experiment, we needed a group of subjects ("users"), an interactive retrieval system with the ability to present results in relevance-ranked lists or in organized clusters, and a design that would insure, as much as possible, that the essential variables in performance would be due to user judgments of documents.

For subjects, we enlisted eight members of the CLARITECH staff. We chose only native speakers of English and tried to

avoid people who had participated in past interactive-retrieval experiments using the CLARIT system. (In fact, only one of the seven subjects had had previous experience using the system.) Among the users were three of the authors of this paper, two other CLARIT developers, and three non-technical volunteers.

For an interactive retrieval system, we chose a version of the CLARIT system that supports both conventional presentation of ranked retrieval results and also automatically clustered results. Since the system has many parameters, we selected a default set and held them constant across all subsequent experiments.

All fifty ad-hoc queries were prepared in advance by two members of the CLARITECH research staff. This was designed to insure that all users would begin their interactions with the same initial queries and that no variability in results would be due to the relative skill (or lack of skill) that individual subjects might have in formulating queries. (We should note that, in the CLARIT system, initial query formulation is typically based on a natural-language statement of the topic and the optional addition of one or more global "constraints" on individual terms. In practice, query formulation is a quick and easy step.)

The 50 ad-hoc topics were randomly[1] divided into six sets of 7 topics and two sets of 4; each user was assigned two sets of topics. (Two users participated only half-time, using the smaller sets.) For one topic set, the user viewed query results in a simple ranked list; for the other set, the top 150 documents retrieved were clustered using CLARIT clustering techniques, and the user was presented with the clusters. Half of the users worked on ranked documents first, while the other half worked on clusters first. Each query was addressed once in each mode, by two different users.

For each topic, users began their interactions by being presented with the initial query and the corresponding initial search results, in ranked or clustered format. Users were instructed to identify as many relevant documents per topic as they could find in 30 minutes, and, along the way, to mark any non-relevant documents that could be useful for negative feedback. For the first 15 minutes, interaction was restricted to review of retrieved documents. Users could scan the terms characterizing a cluster (clustering runs only), read the titles of retrieved documents, or view document text or automatically-generated document summaries to assess document relevance. For the second 15 minutes, users were also permitted to modify the initial query or formulate a new query for the topic. They could reweight query terms, add or delete query terms, modify query constraints, or incorporate system-assisted query enhancements. (This general flow of processing is illustrated in Figure 1.) Judgments were saved automatically at 10 minutes, 15 minutes, 20 minutes, and 30 minutes. User-modified queries were saved at 20 minutes and at 30 minutes.

Subsequent processing of results was fully automatic. We used the accrued judgments collected at the 30-minute point for relevance feedback in the final step of processing over the full TREC target corpus. In each case, we used Rocchio scoring to rank and select 250 "positive" terms from the marked-relevant documents and 15 "negative" terms from the marked-non-relevant documents to supplement the version of the query as formulated at the 30-minute point. This final query was used to retrieve and rank 1,000 documents. In the final submission, we automatically re-sorted the retrieved documents to insure that all previously identified relevant documents were ordered first, followed by the remaining ranked results. We prepared three TREC-7 manual ad-hoc submissions, as follows. (1) A combined set (CLARIT98COMB) based on the unique union of relevance judgments from each mode. (If a document was judged as relevant by one user and as non-relevant by another user, we treated it as relevant.) (2) A cluster set (CLARIT98CLUS) based on the results from the cluster mode. (3) A baseline set (CLARIT98RANK) based on the baseline (relevance-ranked list) mode.

3 General User Performance

From the point of view of general performance, users who interacted with the system in cluster mode rendered more "positive" relevance judgments than users who interacted in ranked mode. In particular, as shown in Table 1, users who interacted in ranked mode recorded 936 positive judgments and 1,626 negative judgments for the 50 topics. Users who interacted in cluster mode recorded 1,025 positive and 1,494 negative judgments.

We did a paired T-test on the numbers of documents users marked as relevant in the ranked sessions versus the number they marked relevant in the cluster sessions, for each time point (10 minutes, 15 minutes, 20 minutes, 30 minutes). At the 15-minute point, in particular, the average number of documents marked should be precisely comparable if clustering has no effect. We found that the cluster sessions runs have higher averages at every time point, although the differences are not statistically significant (see Table 2).

We also regressed the baseline (ranked document) sessions, and found a slope of +7.2. The cluster sessions regressed to a line with a slope of +7.9. A higher slope for clustering could mean that clustering enables the user to find relevant documents faster; however, once again the difference was not significant. We plotted the (*cluster – baseline*) difference points and regressed those data; the slope should be 0 if there is no benefit to clustering. The slope is +1.0—positive but not significant.

[1] The topic sets were randomly generated, except for a restriction that four of the sets contain only topics whose initial queries were written by the same CLARIT researcher. This was necessary because the two researchers who wrote the queries both participated as experimental subjects. To avoid bias, each one worked only with queries written by the other.

4 Retrieval Performance

The official TREC results represent one measure of the relative effectiveness of the two modes of interaction that users engaged in. Though users in the cluster mode submitted more documents to the system as candidate relevants, it was possible that their judgments were inaccurate and that the greater number of documents they nominated would lead to a degradation in system performance. As can be seen in the official results as summarized in Table 3, however, that was not the case. The cluster runs outperformed the ranked ones on all measures, in particular, on both average precision and total recall.

We note also in Table 3 that the overall performance of CLARIT98COMB was superior to that of CLARIT98CLUS on all measures except initial precision. The differences between CLARIT98RANK and CLARIT98COMB are statistically significant at 95% confidence, so we can conclude that clustering has a positive incremental effect. Table 4 provides further information relative to the performance of the three runs. We note again the superior performance, especially in front-end precision, of the CLARIT98CLUS run.

In terms of TREC-group performance, all three submissions performed well above average, as can be seen in Table 5. CLARIT98COMB was below median for only seven topics; and, interestingly, CLARIT98CLUS scored seven “bests”.

5 Effects of Relevance Judgments

In our post-TREC experiments, we compared the NIST judges' and CLARIT users' relevance judgments, and evaluated the relative impact of judgment differences on retrieval performance.

Tables 6–8 summarize the differences between NIST and CLARIT relevance judgments for the documents that were judged by both the NIST judges and CLARIT users for the 50 topics. Table 6 shows the CLARIT user judgments from the ranked run, Table 7 from the cluster run, and Table 8 from the "combined" run, in which we merged the judgments from the baseline and cluster runs for the automatic relevance-feedback retrieval step.

The agreement between the two judgments is calculated by dividing the *number with same judgment* by the *total number of judged documents.*

For the ranked run, agreement is (680 + 1076) / (680 + 1076 + 204 + 256), or 0.7924. For the clustered run, agreement is (703 + 984) / (703 + 984 + 177 + 322), or 0.7717. For the combined run, in which we artificially "lowered" the criterion for "yes" by taking an "OR" operation when merging, agreement is (983 + 1691) / (983 + 1691 + 227 + 512), or 0.7835. This level of agreement is comparable to the levels reported by NIST in studies of inter-rater reliability among TREC judges.

We conducted two sets of experiments to test the effect of the difference in relevance judgments on retrieval performance. We compared our official CLARIT ad hoc runs with the results of experiments that use two different document relevance assessments: first, the "corrected" relevance judgments of the CLARIT users and, second, the relevance judgments of the NIST judges.

Experiments RANKCOR, CLUSCOR, and COMBCOR use the "corrected" relevance judgments of the CLARIT users. CLARIT users' relevance judgments were revised to reflect the relevance judgments of the NIST judges wherever there was a conflict; in cases where CLARIT users judged a particular document but NIST judges did not, the CLARIT user judgment was retained. The resulting numbers of relevant and non-relevant judgments used in the batch feedback step are shown in Table 9.

In all other processing, the RANKCOR, CLUSCOR, and COMBCOR runs were identical to CLARIT98RANK, CLARIT98CLUS, and CLARIT98COMB, respectively. Comparison of retrieval performance is given in Table 10.

In general, we can see that "corrected" relevance judgments have a dramatic impact on the performance of the system. The improvement in recall when judgments are "corrected" is statistically significant at 95% confidence for the ranked-document run; the larger improvement in average precision and precision at 100 documents is statistically significant for all three runs. Of course, from the point of view of the CLARIT users, all the documents they marked as relevant were "correct". Thus, depending on one's point of view, this evaluation can be regarded as giving either a practical upper limit on the performance of the system (the results with NIST judgments substituted selectively for CLARIT user judgments) or a measure of the distortion introduced by conflicting user judgments, which cannot be avoided in actual retrieval applications.

The second set of experiments compares the effectiveness of the relevance feedback based on complete NIST relevance judgments on the set of documents judged by CLARIT users. In RANKNIST, CLUSNIST, and COMBNIST, CLARIT users' relevance judgments were revised to reflect the NIST judges' relevance judgments in all cases, including no judgment if the document was not judged by NIST. The resulting numbers of positive and negative judgments are shown in Table 11.

Note that the numbers of positive judgments for the ranked, cluster, and combined runs are exactly the same when CLARIT judgments are "corrected" (Table 9) and when NIST judgments are substituted (Table 11). This reflects the fact that the NIST judges concentrated on documents that one or more TREC contestants judged relevant; there simply are no documents that CLARIT users judged relevant and NIST judges left unjudged. The numbers of negative judgments do differ from the "corrected" runs to the NIST-only runs, however. The results of the NIST-only run are shown in Table 12. The slight differences between RANKCOR, CLUSCOR and COMBCOR in Table 10 and RANKNIST, CLUSNIST, and COMBNIST in Table 12 must be attributed

to the change in the number of non-relevant judgments included in the batch feedback step.

6 Effects of Timing

Since our TREC experiment automatically saved CLARIT user judgments at 10-, 15- , and 20-minute points, as well as at the end of each 30-minute session, we were able to compare numbers of documents judged at timed intervals (Table 2), and to use these sets of judgments in the batch feedback step in additional experiments (Table 13).

The difference between the ranked and cluster runs was statistically significant at 95% confidence for both recall and precision at 20 minutes; it was not significant at 10, 15, or (the official) 30 minutes.

.

7 Conclusion

We consider our experiments to be a first step in the direction of assessing the effects of information organization on user and system performance. We consider such effects to be critical, especially in a system such as CLARIT which already supports the user in the efficient discovery of relevant information and performs extremely well with a relatively small amount of user feedback.

These first experiments are somewhat inconclusive. We see that clustered representations of retrieved documents lead to overall better performance, both by the user and by the system, but the magnitude of the improved performance is not statistically significant in all cases. The fact that we see such statistically significant improvement at the 20-minute point is especially encouraging, however, since we would hope to see an impact in shorter, not longer, periods of interaction.

The full assessment of the role of information organization in user/system interactions will require a great deal more research. Even in our current design, there are many obvious questions that bear further investigation.

As one example, it would be interesting to know whether the initial queries we used in our experiment were “too good”. One of our hypotheses is that proper clustering and results organization will assist users by concentrating related relevant documents (segregated from non-relevant ones) among a large set of retrieved results. Such an effect would tend to be especially dramatic in the case of a poor or limited initial query, since the relevant documents that respond to such a query are likely not to be serially adjacent to one another—or at the top—in a ranked list. Thus, any process that identifies similar documents and groups them might well succeed in isolating the few relevant documents that respond to a poor query, giving the user an opportunity to identify them more easily. We intend to rerun our experiments with new subjects and impoverished queries to test this hypothesis.

In particular, we intend to repeat the experiments with sparser initial queries, not only to determine whether the original initial queries were "too good" and thus dampened the positive effects of clustering that we observed, but also to determine whether there is a "lower bound" on the effect—a minimal query such that no difference in effectiveness can be observed.

As another example, it will be important for us to experiment with some of the many parameters that exist in our system for clustering, to assess their effect on user performance. Can users work more efficiently with larger or with smaller clusters? Should clusters be summarized via terms or discursively? Should documents within clusters be ranked or organized further with respect to the initial query? There are many such issues that we are only beginning to investigate.

References

1. [Milic-Frayling et al. 1998] Milic-Frayling, Natasa, Chengxiang Zhai, Xiang Tong, Peter Jansen, and David A. Evans, "Experiments in Query Optimization, the CLARIT System TREC-6 Report". In Voorhees, E.M., and Harman, D.K. (Editors), *The Sixth Text REtrieval Conference (TREC-6*). NIST Special Publication 500-240. Washington, DC: U.S. Government Printing Office, 1998, 415–454.

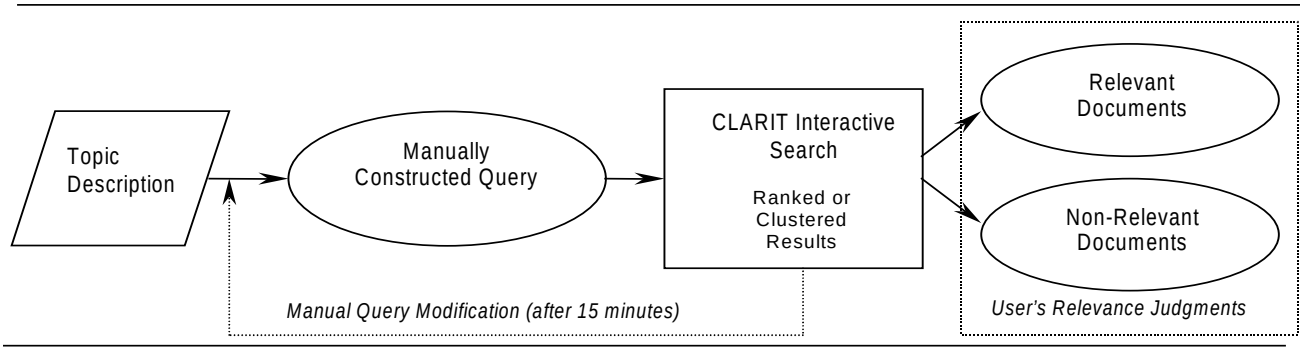

Figure 1. Interactive query formulation and relevance assessment.

Positive judgments	**Ranked session**	**Cluster session**
Total	936	1025
Average	18.72	20.50
Maximum number for one topic	59	75
Minimum number for one topic	1	2
Number of topics w/o positive judgments	0	0
Negative judgments	**Ranked session**	**Cluster session**
Total	1626	1494
Average	32.52	29.88
Maximum number for one topic	116	140
Minimum number for one topic	3	1
Number of topics w/o negative judgments	2	6

Table 1. Statistics about CLARIT users relevance judgments.

	10 minutes	**15 minutes**	**20 minutes**	**30 minutes (basis for TREC runs)**
Ranked runs				
Positive judgments	433	591	693	936
Negative judgments	514	910	1187	1626
Cluster runs				
Positive judgments	457	629	797	1025
Negative judgments	514	847	1080	1494

Table 2. Number of positive and negative judgments in each mode at timed intervals.

Run	**Recall**	**Avg. Precision**	**Initial Precision**	**Exact Precision**	**Prec. 100 docs**
1. CLARIT98RANK	3198	0.3351	0.8814	0.3726	0.2864
2. CLARIT98CLUS	3310	0.3525	0.9066	0.3730	0.2982
(over (1))	(+ 3.50%)	(+5.19%)	(+2.86%)	(+0.11%)	(+0.63%)
3. CLARIT98COMB	3417	0.3702	0.8796	0.4140	0.3178
(over (1))	(+6.85%)	(+10.47%)	(-2.98%)	(+11.11%)	(+10.96%)
(over (2))	(+3.23%)	(+5.02%)	(-0.20%)	(+10.99%)	(+6.57%)

Table 3. Performance statistics for CLARIT98RANK, CLARIT98CLUS, and CLARIT98COMB.

	CLARIT98RANK	**CLARIT98CLUS**	**CLARIT98COMB**
At 5 docs	0.6720	0.7600	0.6920
At 10 docs	0.6440	0.6940	0.6940
At 15 docs	0.6320	0.6360	0.6613
At 20 docs	0.6050	0.5870	0.6180
At 30 docs	0.5327	0.5100	0.5653
At 100 docs	0.2864	0.2982	0.3178
At 200 docs	0.1975	0.1979	0.2142
At 500 docs	0.1074	0.1106	0.1147
At 1000 docs	0.0638	0.0662	0.0683
Exact Precision	0.3726	0.3730	0.4140

Table 4. Precision at N retrieved documents for CLARIT98RANK, CLARIT98CLUS, and CLARIT98COMB.

Run	**Average Precision**			
	>= median	**< median**	**= best**	**= worst**
CLARIT98RANK	33	17	4	0
CLARIT98CLUS	38	12	7	0
CLARIT98COMB	43	7	4	0

Table 5. CLARIT ad hoc results compared to TREC group performance.

		CLARIT Ranked Run		
		Yes	**No**	***Total***
NIST	**Yes**	680	204	884
	No	256	1076	1332
Total		936	1280	2216

Table 6. Comparison of CLARIT user judgments on ranked run with NIST judgments for the same documents.

		CLARIT Cluster Run		
		Yes	**No**	***Total***
NIST	**Yes**	703	177	880
	No	322	984	1306
	Total	1025	1161	2186

Table 7. Comparison of CLARIT user judgments on cluster run with NIST judgments for the same documents.

		CLARIT Combined Run		
		Yes	**No**	***Total***
NIST	**Yes**	983	227	1210
	No	512	1691	2203
	Total	1495	1918	3413

Table 8. Comparison of merged CLARIT user judgments with NIST judgments for the same documents.

	CLARIT98RANK	RANKCOR	*CLARIT98CLUS*	CLUSCOR	*CLARIT98COMB*	COMBCOR
Relevant	*936*	884	*1025*	880	*1495*	1210
Non-relevant	*1626*	1678	*1494*	1639	*2542*	2827

Table 9. Positive and negative judgments after "correction" of CLARIT judgments (official TREC runs shown for comparison, in italics).

Run	Recall	Average Precision	Initial Precision	Exact Precision	Prec. at 100 docs
CLARIT98RANK	**3198**	**0.3351**	**0.8814**	**0.3726**	**0.2864**
RANKCOR	3238	0.4118	0.9828	0.4192	0.3006
(over above)	(+1.25%)	(+22.90%)	(+11.51%)	(+12.51%)	(+4.95%)
CLARIT98CLUST	**3310**	**0.3525**	**0.9066**	**0.3730**	**0.2982**
CLUSCOR	3316	0.4165	1.0000	0.4193	0.3062
(over above)	(+0.18%)	(+18.17%)	(+10.30%)	(+12.41%)	(+2.68%)
CLARIT98COMB	**3417**	**0.3702**	**0.8796**	**0.4140**	**0.3178**
COMBCOR	3410	0.4579	0.9806	0.4550	0.3254
(over above)	(-0.20%)	(+23.70%)	(+11.48%)	(+9.92%)	(+2.39%)

Table 10. Effects of user feedback based on CLARIT users' judgments and "corrected" relevance judgments.

	CLARIT98RANK	RANKNIST	*CLARIT98CLUS*	CLUSNIST	*CLARIT98COMB*	COMBNIST
Pos. judg.	*936*	884	*1025*	880	*1495*	1210
Neg. judg.	*1626*	1332	*1494*	1306	*2542*	2203

Table 11. Positive and negative judgments using NIST judgments only (official TREC runs shown for comparison, in italics).

Run	Recall	Average Precision	Initial Precision	Exact Precision	Prec. at 100 docs
CLARIT98RANK	3198	0.3351	0.8814	0.3726	0.2864
RANKNIST	3236	0.4114	0.9829	0.4188	0.3006
CLARIT98CLUST	3310	0.3525	0.9066	0.3730	0.2982
CLUSNIST	3312	0.4164	1.0000	0.4196	0.3060
CLARIT98COMB	3417	0.3702	0.8796	0.4140	0.3178
COMBNIST	3406	0.4577	0.9806	0.4552	0.3258

Table 12. Effects of user feedback based on NIST judgments only.

Ranked run	Recall	Average Precision	Initial Precision	Exact Precision	Prec. at 100 docs
10 min	3036	0.2806	0.8635	0.3228	0.2488
15 min	2994	0.3007	0.8729	0.3400	0.2608
20 min	3017	0.2982	0.8734	0.3416	0.2638
30 min	3198	0.3351	0.8814	0.3726	0.2869
Cluster Run					
10 min	3060	0.2847	0.8430	0.3189	0.2468
15 min	3141	0.3099	0.8507	0.3426	0.2766
20 min	3171	0.3324	0.8773	0.3616	0.2848
30 min	3310	0.3525	0.9066	0.3730	0.2982

Table 13. Results using judgments at timed intervals, ranked and cluster runs.