TREC-2004 HARD-Track Experiments in Clustering

David A. Evans, Jeffrey Bennett, Jesse Montgomery, Victor Sheftel,
David A. Hull, James G. Shanahan

Clairvoyance Corporation, Pittsburgh, PA
{dae; bennett; montgomery; sheftel; hull; jimi}@clairvoyancecorp.com

1. Introduction

The Clairvoyance team participated in the High Accuracy Retrieval from Documents (HARD) Track of TREC 2004, submitting three runs. The principal hypothesis we have been pursuing is that small numbers of documents in clusters can provide a better basis for relevance feedback than ranked lists or, alternatively, than top-*N* pseudo-relevance feedback (PRF). Clustering of a query response can yield one or more groups of documents in which there are "significant" numbers (greater than 30%) of relevant documents; we expect the best results when such groups are selected for feedback. Following up on work we began in our TREC-2003 HARD-Track experiments [Shanahan et al. 2004], therefore, we continued to explore approaches to clustering query response sets to concentrate relevant documents, with the goal of (a) providing users (assessors) with better sets of documents to judge and (b) making the choices among sets easier to evaluate. Our experiments, thus, focused primarily on exploiting assessor feedback through clarification forms for query expansion and largely ignored other features of the documents or metadata.

One of the submitted runs was a baseline automatic ad-hoc retrieval run. It used pseudo-relevance feedback, but no assessor judgments. Two non-baseline runs contrasted alternative strategies for clustering documents in the response set of a topic—one based on simple re-grouping of responding documents (our "standard" approach using quintad pseudo-clusters) and another based on reprocessing of the response set into small sub-documents (passages) and then clustering. In both cases, the grouped documents were presented to assessors in evaluation forms and the groups of documents selected as being on topic were used as the source of query expansion terms. A third version of the response set, based on summaries of sub-document clusters, was also submitted to assessors, but not as an official run.

Our standard approach, using simple re-grouping of documents into quintad pseudo-clusters, proved robust and most effective, giving above-median performance. However, our attempt at producing more concentrated clusters of relevant information by using small sub-documents instead of whole documents in clusters proved to be largely unsuccessful. Upon further analysis, this result seems to derive more from the apparent difficulty assessors may have had in judging such sub-document clusters and less from the actual quality of the clusters or our ranking of them. We attribute this failure to our inability to discovery the best methods (parameters) for clustering a response set and to the limitations in representing the results of response-set analysis through the rather rigid forms interface to assessors.

In the following sections, we provide the details of our work. Section 2 presents our processing approach and experiments, including examples of the assessor forms we produced. Section 3 gives a brief summary of our results. Section 4 offers an analysis of our work from the point of view of our primary hypotheses. Section 5 gives concluding thoughts.

2. Experiments

All Clairvoyance processing for the HARD-Track experiments was managed under the CLARIT system, thus, terms (features) were based on morphologically-normalized linguistic phrases for document indexing and query processing. Furthermore, all documents were processed as sets of sub-documents and the scoring of documents was based on best-scoring sub-documents. (Cf. [Evans & Lefferts 1995] and [Milic-Frayling et al. 1997] for representative descriptions of CLARIT processing.)

To form a query for a topic we merged the topic's title, description, and narrative (TDN) fields. We extracted terms from the query text using CLARIT natural-language processing, keeping all normalized simplex noun phrases (NPs), attested sub-phrases, and contained single words as terms [Evans & Lefferts 1995]. We also included all adjectives and adverbs that appeared outside noun phrases. Since the same units—NPs, sub-phrases, words, etc.—are used to index documents (sub-documents), the terms (features) extracted from the query text are commensurable with those from the corpus.

Each resulting query term, *t*, is associated with a weight that is computed as follows:

$$Weight(t) = TF(t) \times IDF(t) \times coefficient(t) \quad (1)$$

where the *coefficient(t)* value is set to *1* and *TF*, the term frequency, is defined as follows:

$$TF(t) = 0.5 + 0.5 \times TermFreq(t) \quad (2)$$

where *TermFreq(t)* denotes the number of times the term *t* occurs in the query.

The *IDF* term, corresponding to the inverse document frequency, is defined as follows:

$$IDF(t) = 1 + \log\left(\frac{SubDocCount}{SubDocCount_t}\right) \quad (3)$$

Where *SubDocCount* is the number of sub-documents in the corpus and $SubDocCount_t$ corresponds to the number of sub-documents in the corpus that contain the term *t*.

In the case of HARD-Track experiments, we merged all HARD-Track test corpora into a single large corpus over which global statistics (such as *IDF*) were computed. In the CLARIT system, sub-documents are defined at indexing time. The sub-document size is fully configurable, with the default setting for our experiments producing passages on paragraph boundaries that range in size from 8-to-20 sentences. The default sub-document size is 12 sentences if paragraph boundaries are not found within 20 sentences.

Queries submitted to the CLARIT retrieval engine result in a ranked list of sub-documents based upon the score between each sub-document (*SubDoc*) and the query. This sub-document score is calculated as follows:

$$Score(SubDoc,Query) = \sum_{t \in Query} TF_{SubDoc}(t) \times IDF(t)^2 \times coefficient(t) \times TF_{Query}(t) \quad (4)$$

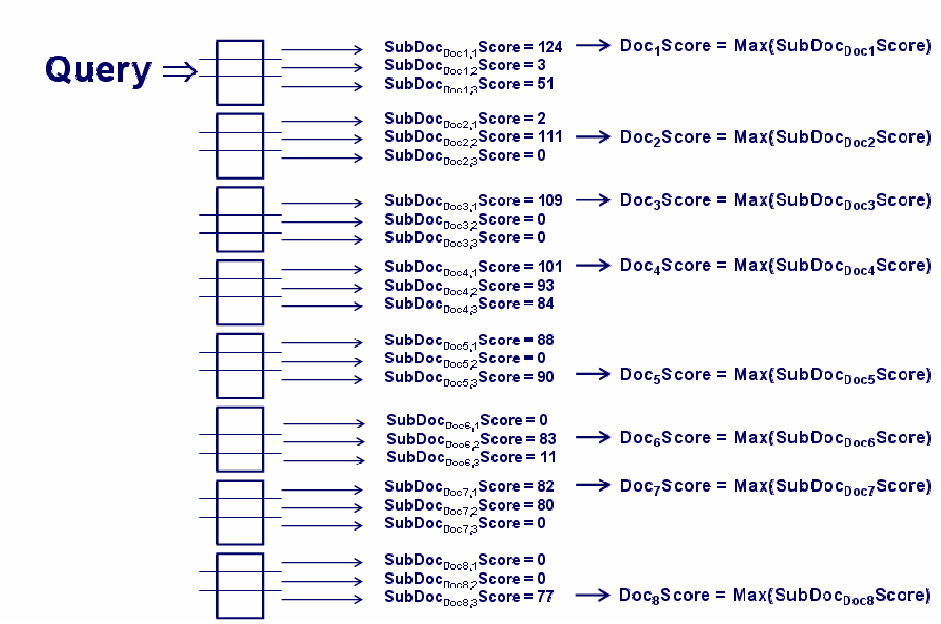

Figure 1. Schematic Representation of Document Scoring Based on Max-Sub-Doc Score

The ranked list of sub-documents is post-processed such that sub-documents belonging to a single document are reassigned to the original document and its (document) score is set to the score of the highest scoring sub-document. This is illustrated schematically in Figure 1.

Subsequently, pseudo-relevance or blind feedback can be used to expand the original query automatically using terms extracted from the top *C* ranked documents. PRF has been shown to improved ah-hoc retrieval performance. (Cf. [Evans & Lefferts 1994] for a description of the first such effects of pseudo-relevance feedback in the context of TREC document processing.) A similar process can be used for supervised or directed feedback. Terms are extracted from all sub-documents that score at or above the *C*-th document score. This may lead to selection of multiple sub-documents from the same document, as illustrated in Figure 2 for "top-6" document PFR.

Extraction of terms ("Thesaurus Extraction") from selected top sub-documents is performed using CLARIT NLP. Terms are selected for feedback (merging with the query) via the following steps. Terms are ranked in decreasing order using the *Prob2* weighting scheme, which is defined as follows:

$$Prob2(t) = \log(R_t + 1) \; x \left(\log(\frac{N - R + 2}{N_t - R_t + 1} - 1) - \log(\frac{R + 1}{R_t} - 1) \right) \quad (5)$$

where N is the number of sub-documents in the reference corpus, and N_t is the distribution of t in the corpus (i.e., the number of sub-documents that contain the term t in the corpus) . Similarly, R is the number of sub-documents in the top C documents, and R_t is the distribution of the term in the top C documents (i.e., the number of sub-documents that contain the term t in the top C documents). The top k terms (highest *Prob2*-weighted terms), known as the expanded set, are merged with the original query. In all our experiments, we set C to *10* (documents) and k to *200* (terms).

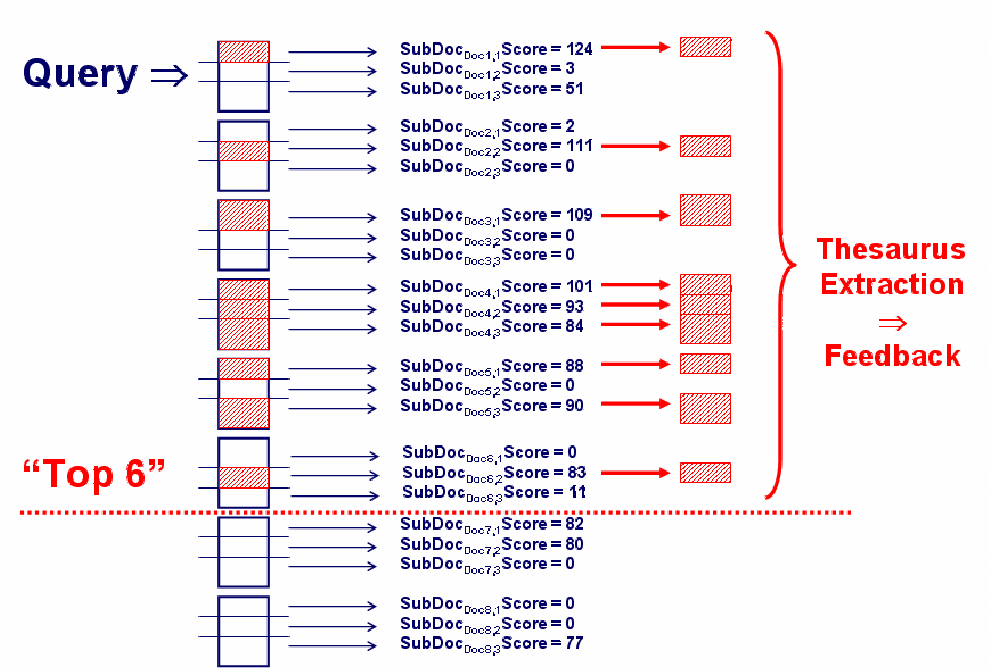

Figure 2. Schematic Representation of Sub-Doc Selection for "Top-6" Document Feedback

We set the coefficients of query terms and new terms (*coefficient(t)* in Formula 1) through a term normalization algorithm. The coefficient of a term that occurs in both the expanded set and in the original query is set to *(1 x boostFactor) + normalisedProb2*. The coefficient of a term that occurs in the query only is set to *1.0*. Finally, the coefficient of a term that occurs only in the expanded set is set to *normalisedProb2*. For our experiments, *boostFactor* was *2*. The *normalisedProb2* factor is calculated as follows:

$$normalisedProb2(t) = \frac{Prob2(t)}{MaxProb2} \qquad (6)$$

where *MaxProb2* is the highest *Prob2* score for any term in the expanded set under consideration.

2.1. Baseline Run (CL102TDN)

Our baseline submission (CL102TDN) reflected an automatic pseudo-relevance feedback run. Queries were formed from the title, description, and narrative (TDN) portions of each topic and then submitted to the full test corpus. Two hundred terms were selected from the top-responding 10 documents for each query based on CLARIT *Prob2* term scoring. (Operationally, as noted above, it is the highest-scoring sub-documents in the top-10 documents that are used for feedback, not the whole documents.) These terms were given weight-normalized coefficients (as described above) and the terms selected for feedback were merged with the terms in the initial query. Any initial-query terms that reappeared in the feedback term set were given coefficient boosting (as described above). The expanded query was resubmitted to the test corpus to obtain the principal results set. No assessor forms were produced for this run.

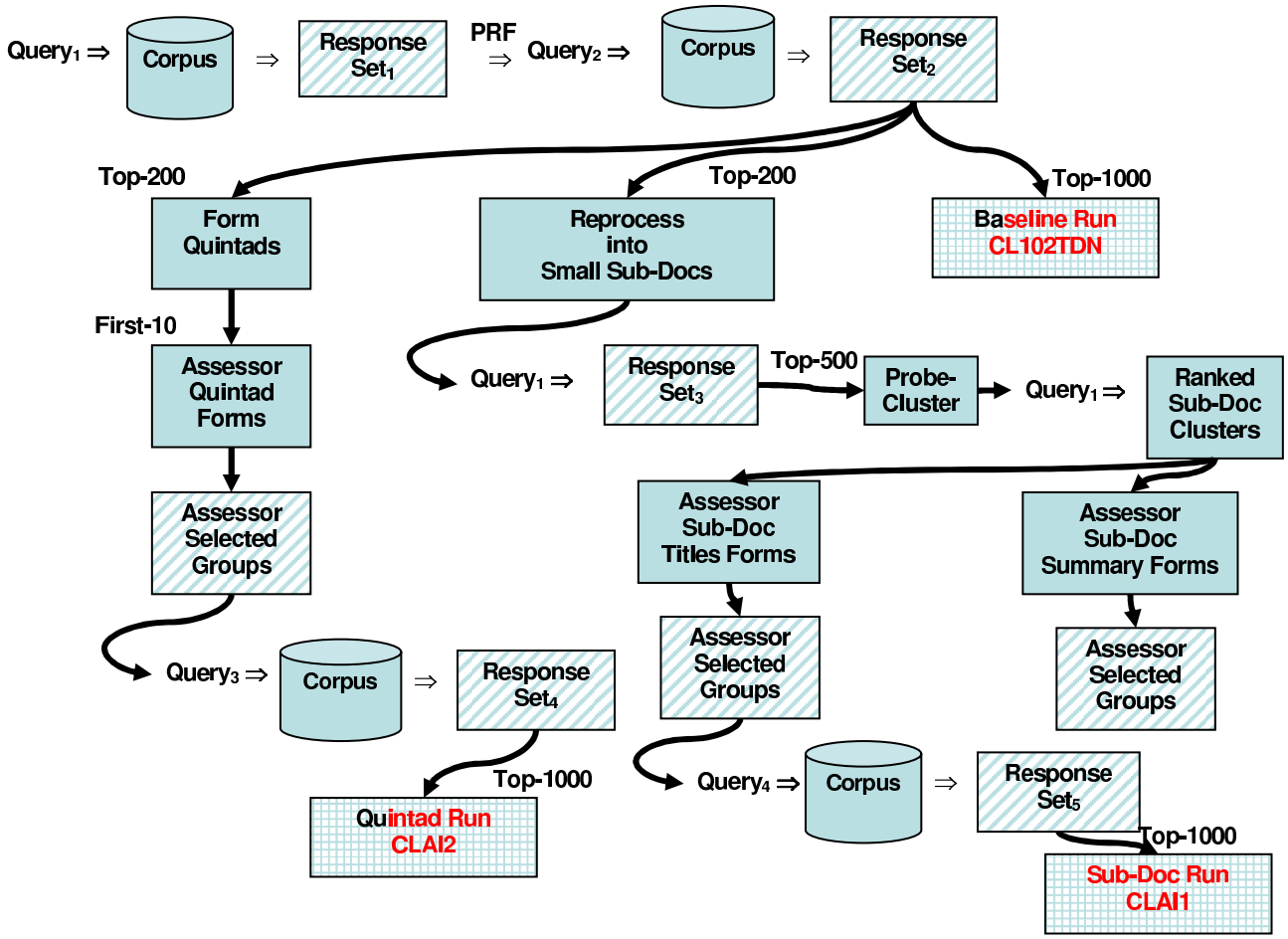

Figure 3. Schematic Representation of Processing for Each Submitted Run

A schematic illustration of the processing involved in producing this baseline run is given in Figure 3. This figure also shows the processing associated with the preparation of the three assessor forms, based on quintads, sub-doc titles, and sub-doc summaries, and the derivation of the other two submitted runs, CLAI1 and CLAI2, as described in the following sections.

2.2. Document-Clustering Run (CLAI2)

As a "standard" experimental run (CLAI2), we used a version of our "quintad" approach of TREC 2003. (Cf. [Shanahan et al. 2004].) Given the ranked results of the baseline run, the top-200 ranked documents for each query were segregated as a response set. For each topic, we automatically organized the response set into groups of documents as follows. Beginning with the first ranked document, a query was created from the document by parsing its top-scoring sub-document; it was then submitted to the topic's response set (of 200 documents). The source document and the four other closest documents that responded to this query were combined into a group (or pseudo-cluster) and removed from the response set. This process was repeated using the next remaining highest ranked document in the response set for the next query. This process continued until a total of ten five-document groups was formed. Note that, using this method, the response set is reorganized with a bias toward the initial top-scoring documents.

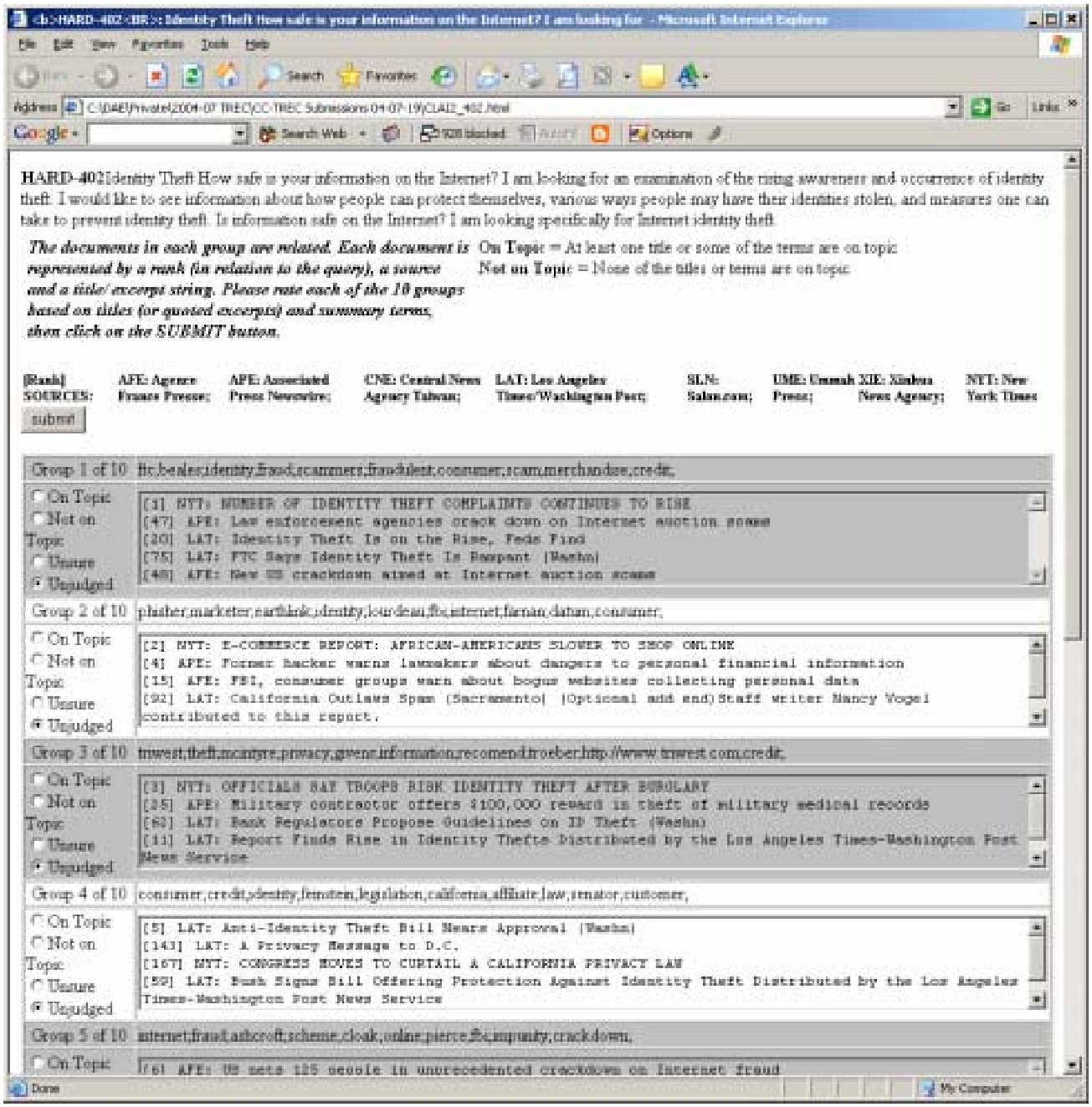

Figure 4. Assessor Form for Topic 402—Document-Clustering (Quintad) Run (CLAI2)

We generated user-feedback (assessor) forms automatically by presenting each of the ten groups in the order in which they were produced. Each group was headed by a "summary" set of 35 terms, taken from the five top-scoring sub-documents in the group (as determined by *Prob2* term scoring), and each group also showed the titles of the documents it contained. (If an included document had no title, a pseudo-title was created by taking the first 100 characters from the beginning of the best-hit sub-document of the document, respecting word boundaries.) A sample assessor form, for Topic 402, is given in Figure 4.

Assessors were asked to judge which groups seemed relevant to the topic and indicated their judgments by selecting the appropriate category: "on topic"; "unsure"; "not on topic"; or "unjudged". The "on topic" option denotes that at least one document or some of the terms in the group are on topic. The "not on topic" option means that the assessor believes none of the group documents or group terms represents the topic well. If uncertain whether the group accurately represents the topic, the assessor can choose "unsure." And if the assessor runs out of time or simply fails to judge a group, the default value for a group is "unjudged."

All best-scoring sub-documents in any quintad that received an assessor-feedback score of "on topic" were reserved for further processing. If there were fewer than two "on topic" quintads, all best-scoring sub-documents from up to five "on topic" and "unsure" marked groups were used. Two hundred terms were extracted from the selected documents, their weights were normalized, and they were merged back into the TDN version of the query, with boosting. This expanded query was submitted to the full test corpus and the top-1,000 ranked results were provided to TREC as the output of this run, with the following modification: all "on topic" marked documents were promoted to the top of the list and all "not on topic" documents were excluded from the list. In those cases when there was no "on topic" or "unsure" marked group, the original response set from the baseline run was used as the basis of the new result. In such cases, we excluded from the baseline results any documents from groups that may have been marked "not on topic."

2.3. Sub-Document-Clustering Run (CLAI1)

As a second experimental run (CLAI1), we used a new approach to response set clustering. As with the quintad run (CLAI2), the baseline run's top-200 ranked documents for each query were segregated as a response set. This set was reprocessed as a database of short sub-documents (averaging 3–5 sentences each). These short sub-documents were re-ranked against the original query and up to the top-500 responding sub-documents were reserved as response set-2. The sub-documents in response set-2 were grouped using asymmetric clustering.

In one version of asymmetric clustering, a document is chosen at random as a "seed"; its top-ranked *n* terms are used as a query over the target set; the top-responding *k* (sub-)documents are taken as a feedback group; *m* terms are extracted from these (sub-)documents and used as a new query over the target set; all responding documents that score above a threshold, θ (typically expressed as a percent of the highest scoring document's score), are segregated as a cluster and removed from the target set. In our experiments, at any iteration step, we chose as the first or next seed the highest ranked sub-document that was available in the corpus (set-2), as scored against the original query. We set parameters as follows: $n = 15$; $k = 3$, $m = 25$, and $\theta = 0.20$.

Unlike the quintad run, where the number of clusters and documents was fixed (10 clusters with 5 documents each), the sub-document-clustering run did not have a pre-determined number of clusters or composition. Excluding singletons, there was no limit on the number of groups or their contained sub-documents, up to the number of sub-documents in the response set (set-2). Practically speaking, however, the number of clusters that resulted from this process for each topic was 5–25. Each cluster was score-ranked against the original query, using the cluster centroid vector as a representation of the cluster content. Assessor-feedback forms were generated by presenting the groups in rank order; each group was preceded by a set of summary terms based on the centroid vector and each showed the top-five document titles corresponding to the best sub-documents in the cluster. Any summary terms that were also query terms were front-ranked and highlighted.

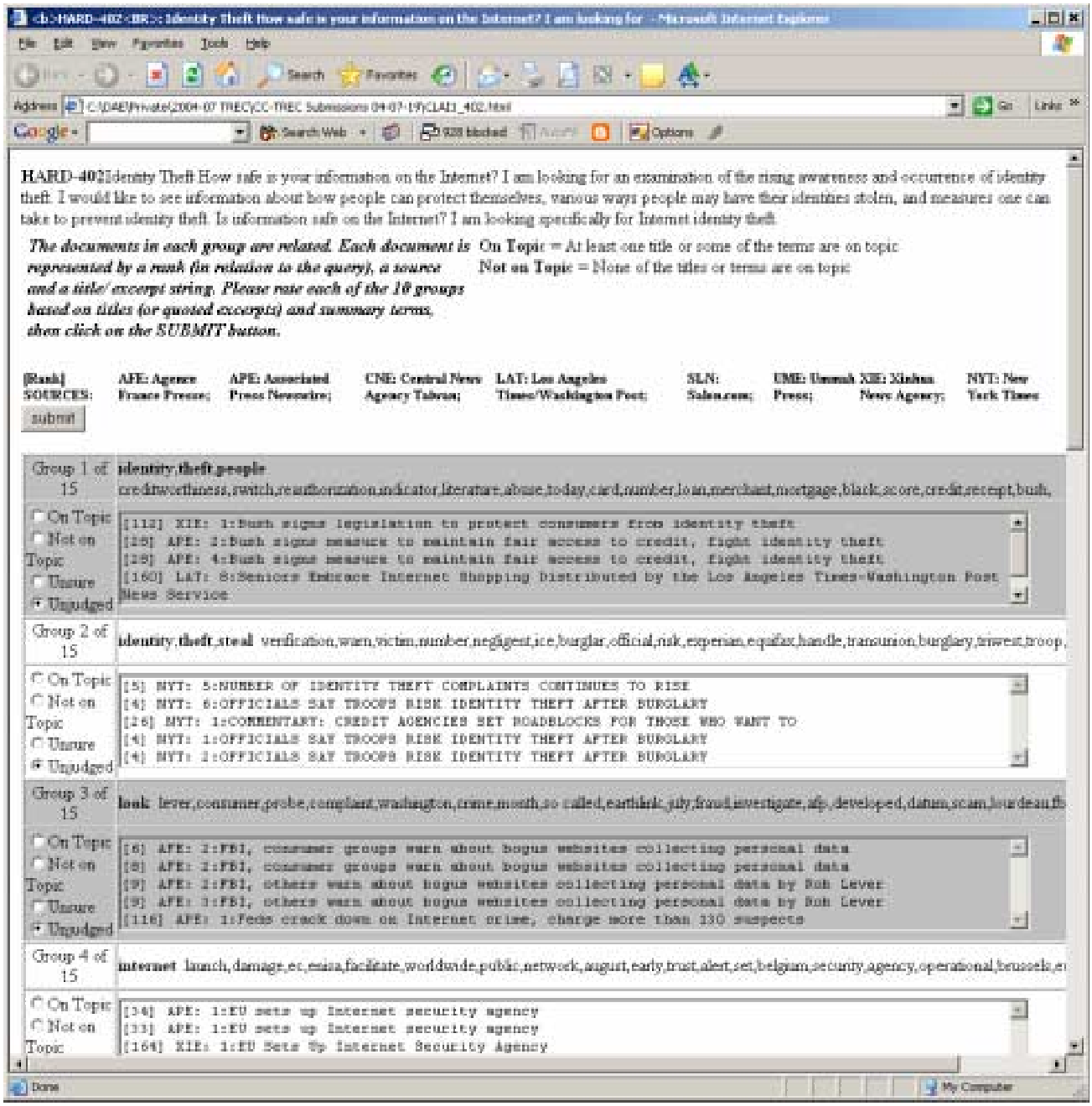

Figure 5. Assessor Form for Topic 402—Sub-Document-Clustering Run (CLAI1)

A sample assessor form for the sub-document-clustering run, for Topic 402, is given in Figure 5. Judges marked groups using the same categories as for the quintad run (CLAI2). Feedback was processed in the same fashion, but was based on the short sub-documents from response set-2.

2.4. Sub-Document-Clustering Summary Run (CLAI3—Unofficial)

As an additional strategy to present information about cluster content to assessors, we created a second set of assessor forms for the sub-document clusters produced for the sub-document

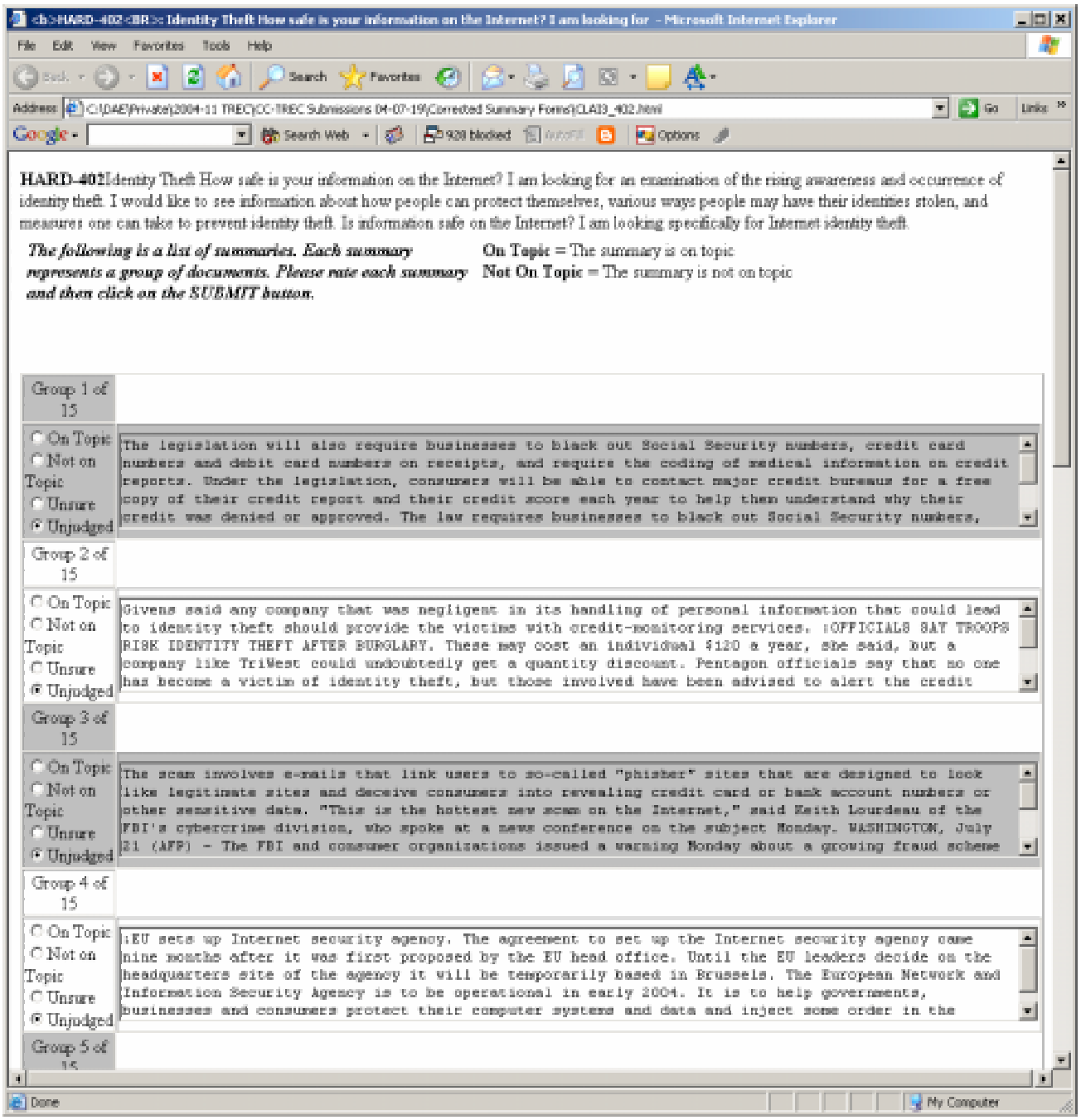

Figure 6. Assessor Form for Topic 402—Sub-Document-Clustering with Summaries (CLAI3)

clustering run (CLAI1). Thus, the actual clusters and their order were identical to the results produced for CLAI1. In this case, however, instead of showing document titles and summary terms, we displayed a textual summary of the clusters. Each cluster summary was based on the whole text of the titles and contents of the five top-scoring sub-documents in the cluster. To create this pseudo-whole document, we simply concatenated the selected titles and sub-documents, allowing repetitions of titles (if required). We then generated a 5-sentence CLARIT summary (a *summary-quintad*) of the pseudo-whole document. A clarification form of this type is shown in Figure 6.

3. Results

Table 1 presents a list of the results for each of our runs and the TREC median and maximum runs, as scored for both the Hard (H) relevance judgments and the Hard and Soft (H&S) judgments. As can be seen in the results, both our baseline run (CL102TDN) and quintad run (CLAI2) scored above median for the track, but significantly below the maximum. The sub-document clustering run was far below median and showed a marked degradation over our baseline and quintad results.

Figure 7 gives the comparative performance results for our three official runs, as scored on individual topics using Hard and Soft relevance judgments. The topics are shown on the x-axis in descending rank order as given by the baseline (CL102TDN) run. In general, the quintad run tracks the baseline run, giving results that are within a range of better than or only slightly worse than the baseline in all but eight instances. In contrast, the sub-document clustering run gives results almost always lower than the baseline (eighteen times much lower) and only seven times significantly higher.

Experiment	H Avg Prec	H R-Prec	H&S Avg Prec	H&S R-Prec
CL102TDN (Baseline)	0.2660	0.2713	0.2613	0.3024
CLAI1 (Sub-Documents)	0.1421	0.1733	0.1716	0.2147
CLAI2 (Quintad)	0.2735	0.2903	0.2872	0.3157
TREC median	0.2617	0.2690	0.2634	0.2906
TREC max	0.3635	0.3766	0.3554	0.3717

Table 1. Mean Average Precision (MAP) Results for Official Runs vs. TREC Median & Max

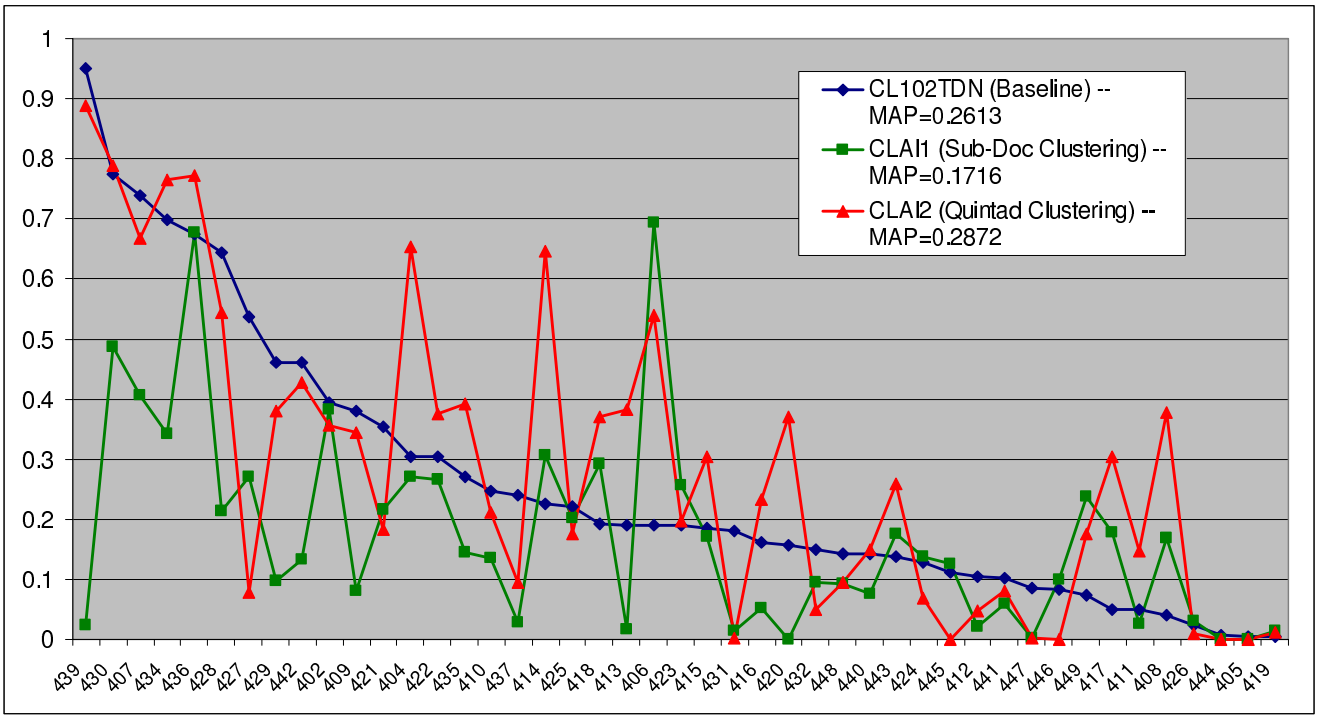

Figure 7. Per-Topic Comparative Performance: CLAI1 and CLAI2 vs. CL102TDN Baseline

4. Analysis and Discussion

Our analysis of the results has focused principally on several questions related to the essential clustering hypothesis. In particular, we hoped to assess whether the clusters we generated, in fact, succeeded in concentrating relevant documents and whether the groups of documents we offered to assessors based on these clusters were identified as relevant. We present our exploration of these questions in the following three sections. Please note that, in all the following assessments, we used the Hard relevance judgments, not Soft.[1]

4.1. Basic Clustering

Did we manage to cluster documents or sub-documents such that relevant documents (or sub-documents) were grouped more densely within the clusters than in the response set as a whole?

Our basic answer is *Yes*, but we demonstrated better success with documents (i.e., in the quintad process) than with sub-documents.

Our analysis is based on the evaluation of the "top" or "best" clusters that resulted from either the quintad or sub-document groupings. In particular, we were interested in seeing whether the clustering process produced one or more clusters that had a density of relevant documents greater than either the response set, generally, or the top portion of the response set, which might be used for a process such as PRF.

The results of this analysis are presented in Figures 8a and 8b (for quintads) and Figures 9a and 9b (for sub-document clusters). Note that we have used only 25 of the available topics for our analysis. In a number of cases, we had no relevance judgments for topics (i.e., there were no relevant documents) or we had no "on topic" assessor judgments. For simplicity and consistency, we chose twenty-five representative topics for which there were both relevant documents and assessor judgments in all response conditions and used these in all phases of the analysis.

For each of the twenty-five topics, ranked in decreasing order of response-set density, we give in left-to-right order the (i) the density of relevant documents in the initial response set, (ii) the density of relevant documents in the top-15 portion of the response set, and (iii) the density of the relevant documents in the (up to) three best groups (quintads or sub-document clusters) that were created by post-response-set processing. In identifying the best groups of documents, we considered both the best groups among the groups that were "visible," i.e., made available to assessors (Figures 8a and 9a), and also the best groups that were created from the response set as a whole, but possibly not chosen for the assessor forms (Figures 8b and 9b).

For each topic given in the figures (graphs), the first two bars show response-set density and top-15 density, respectively. The third (yellow) through fifth (if available) bars show the density in the best-formed document groups. As is apparent from all the figures, clustering by any method does indeed have the effect of concentrating relevant documents. More precisely, clustering over the whole response set (Figures 8b and 9b) shows that, in every case among the twenty-five topics, there is at least one cluster (and often all top three) with relevant document density greater than the response set or the top-15 ranked documents. If we restrict the analysis to the groups that were presented to assessors (Figures 8a and 9a), the effect is still quite strong. Among quintads, only

[1] For the Soft judgments, we only have results for Part 1 of the analysis; the results are not significantly better than for the Hard judgments.

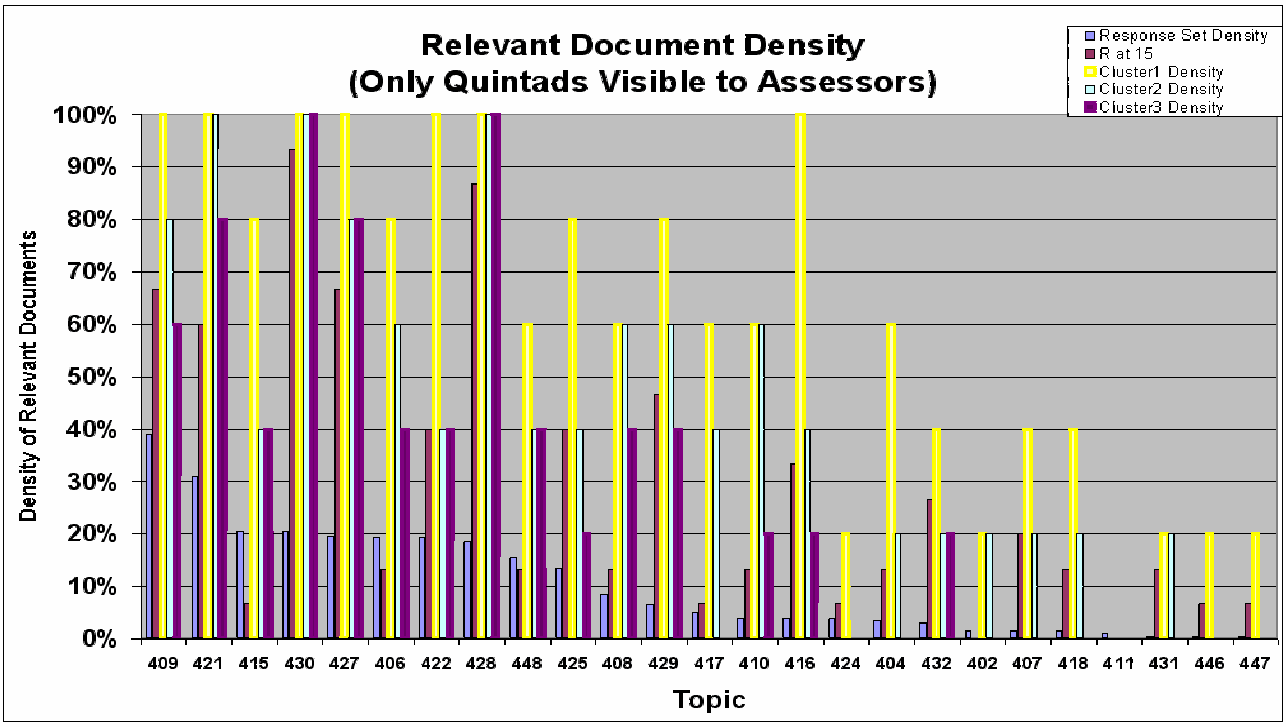

Figure 8a. Density of Relevant Documents in Top Visible Quintads vs. Whole Response Set

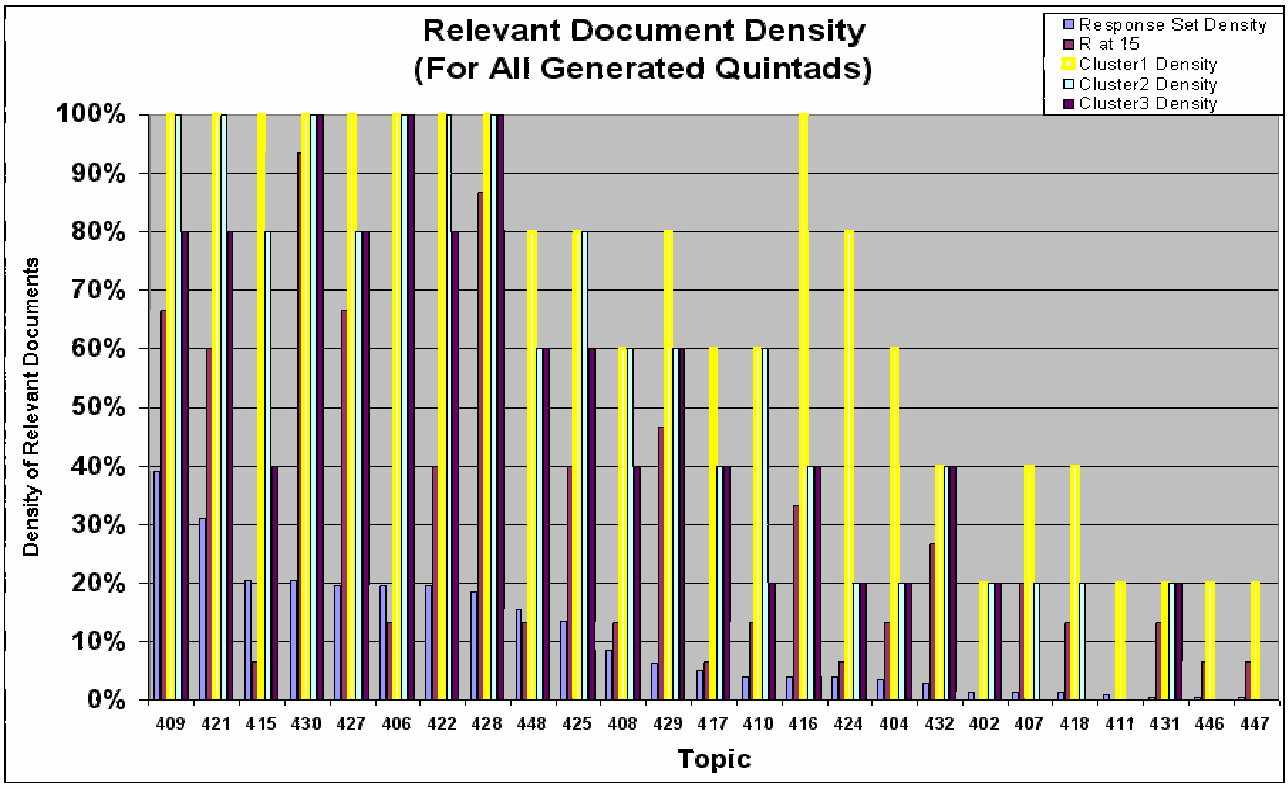

Figure 8b. Density of Relevant Documents in All Top Quintads vs. Whole Response Set

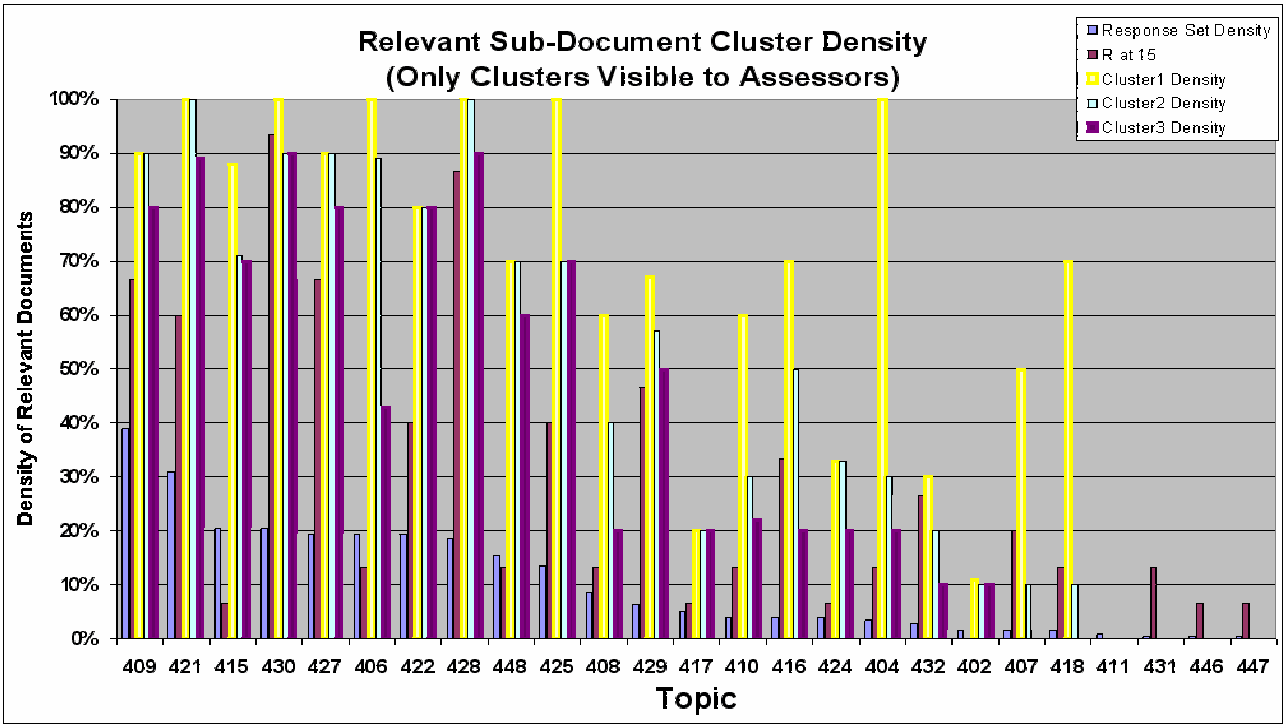

Figure 9a. Density of Relevant Sub-Documents in Visible Clusters vs. Whole Response Set

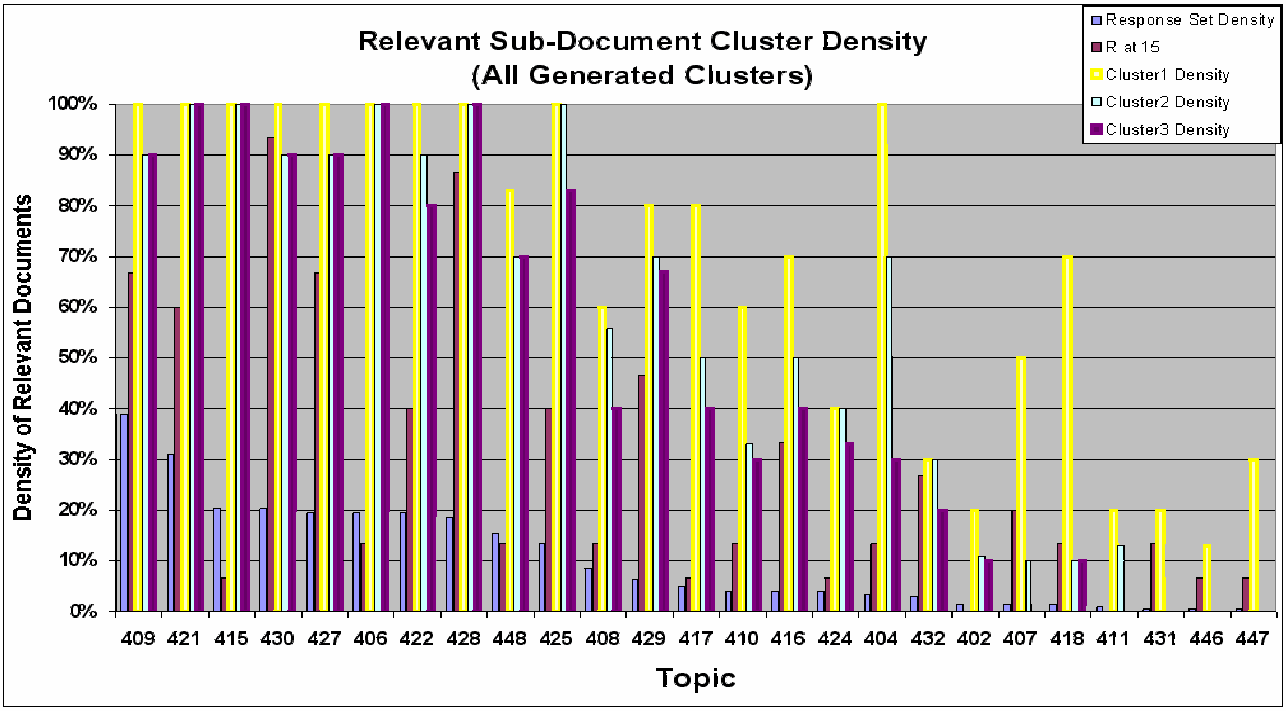

Figure 9b. Density of Relevant Sub-Documents in All Top Clusters vs. Whole Response Set

topic 411 had no groups above background density. Among sub-document clusters, only topics 411, 431, 446, and 447 showed none. More significantly, the vast majority of topics—across the wide range of response-set background densities—showed two or more clusters with significant concentrations of relevant documents (>40%). In short, it appears that post-processing the response set into groups, either via quintad formation or via sub-document clustering, succeeds in isolating and concentrating relevant documents.

4.2. Cluster Selection (Ranking)

Did we manage to select the dense clusters to present to the assessor? In other words, did we rank our clusters correctly?

Our basic answer is *Yes*, but, again, we demonstrated better success with documents than with sub-documents.

We analyzed the document clusters in the rank order we presented them to assessors, as shown in Tables 2 and 3. Topics in the tables are ranked along the x-axis in descending order by MRR score (last row of the tables). The numbers in individual cells represent the density of relevant documents in the clusters; red cells are the densest clusters, and so on. The rows of the table correspond to the cluster rank.

In the case of quintads (Table 2), sixteen of the twenty-five topics have a group with relevants ranked as the top-most cluster. Eighteen topics have such a group at rank one or two. Twenty-two have such a group at rank one, two, or three. This suggests that some concentration of relevants is occurring in the top-most clusters. Coupled with the observations we made about background density in the top-15 documents of the response sets above, this further suggests that taking the first three quintads instead of the top-15 documents might be an alternative, superior strategy in selecting documents for PRF.

However, by other measures the picture is less clear. The "mass" of relevant documents in the groups ranked one to ten (i.e., the groups presented to assessors) is greater than the mass in the balance of the ranked groups in only thirteen cases. In short, not all candidate (good) clusters are being selected and ranked in the top echelon.

In the case of sub-document clusters (Table 3), ten of the topics have groups with relevant documents ranked first. Sixteen topics have such a group at rank one or two; eighteen at rank one, two, or three. This result is poorer than with quintads. One source of variability certainly has to do with the basic clustering strategy. The quintad approach is "conservative": grouping is biased toward initially top-ranked documents. If the original query processing yields some high-ranked relevants, the quintad method has a reasonably good chance of finding (and concentrating) similar documents. In contrast, the sub-document clustering process results in a completely re-analyzed response set and cluster groups must be re-ranked against the initial query in a separate process. It is not clear that the method we have used for scoring (and ranking) clusters—based on matching the query terms with those in the cluster centroid—is appropriate or sufficiently powerful. Compared to scoring documents against a query, for example, where there may be many terms in the document, including many repetitions of the same term, scoring against a cluster centroid is impoverished: our cluster centroids consist of only twenty-five terms and each term is unique.

Ranking of QUINTADS as Presented to Assessors

	421	430	409	427	448	428	415	408	422	425	429	432	416	410	406	407	418	446	447	417	431	404	402	424	411
1	0.40	0.60	0.20	0.80	0.40	0.80	0.40	0.60			0.80	0.20	0.20	0.60		0.40	0.40	0.20	0.20						
2	1.00	1.00	0.60	0.60	0.20	1.00		0.40													0.20	0.60			
3	0.40	1.00	0.80	0.60	0.20	0.60			0.40	0.20			0.40							0.60			0.20		
4	0.40	0.60	0.60	1.00			0.20		0.40	0.20	0.40				0.80									0.20	
5	0.40	0.80		0.60	0.20	1.00	0.80	0.60	0.20	0.80					0.20					0.40					
6	1.00	1.00	0.40	0.40	0.60	0.60				0.20		0.20		0.60											
7	0.80	0.40		0.20			0.40		1.00			0.20	1.00								0.20	0.20			
8	0.20	0.40	1.00	0.40						0.40	0.60			0.20	0.60										
9	0.80	0.20			0.40	1.00	0.40		0.20						0.40	0.20							0.20		
10	0.80	0.60	0.20	0.80					0.20			0.40		0.20			0.20								
11	0.20			0.40	0.20				0.60	0.60	0.60				1.00					0.20					
12	0.80	0.40													0.80										
13	0.40	0.40	0.60	0.60	0.80		1.00		0.80															0.80	
14	0.40	0.40	0.60		0.20		0.20			0.20					1.00					0.20					
15	0.40		0.40			0.40			0.20	0.80	0.20			0.20										0.20	
16			0.60		0.20					0.40					0.80							0.20			
17	0.60		0.40		0.60		0.40													0.40					
18	0.80			0.20	0.20	0.20			0.60																
19	0.20	0.20	0.40	0.40			0.20																0.20		
20	0.40	0.20	0.20										0.40		0.20										
21	0.80						0.40			0.80															
22		0.20	0.20								0.40	0.40													
23	0.80	0.20		0.20		0.20	0.40								0.80						0.20			0.20	
24	0.80								0.20																
25	0.20						0.40			0.60															
26			0.40		0.20		0.20					0.40													
27	0.20						0.40		1.00	0.20															0.20
28			0.40		0.20					0.20			0.40		1.00										
29			0.20		0.20				0.60						0.20										
30	0.20				0.40				0.20		0.20				0.20										
31	0.20		0.60		0.40																		0.20		
32			1.00						0.20	0.20															
33			0.40	0.20	0.20		0.20		0.20	0.40															
34			0.80																						
35	0.40																								
36	0.60	0.20	0.20							0.20		0.40			0.40										
37			0.20		0.20							0.20													
38	0.60		0.20																						
39			0.40		0.20				0.20																
40		0.20	0.60							0.20		0.20									0.20				
MRR	3.89	3.40	3.32	3.17	2.98	2.48	2.20	1.70	1.66	1.64	1.61	1.57	1.56	1.46	1.22	1.11	1.10	1.00	1.00	0.75	0.71	0.71	0.53	0.44	0.04

Cluster Density >= 0.8
0.8 > Cluster Density >= 0.4
0.4 > Cluster Density > 0
Clusters Visible to the Assessor

Table 2. Ranking of Quintad Clusters as Presented to Assessors

In terms of mass of relevants visible to assessors, only eight topics had more relevants available below the threshold of visibility. However, we made fifteen of the sub-document clusters available for assessors, as compared to only ten in the case of quintads. If we limit our analysis of mass distribution to the top-ten sub-document clusters, then seventeen out of twenty-five topics had greater numbers of relevants *below* the threshold. This reinforces the observation above that our approach to ranking is poor.

In general, it is difficult to assess the effectiveness of our ranking or to compare the quintad and sub-document clustering strategies. One source of uncertainty derives from what it means to be "relevant" or to count as a "relevant document" when the units of text that are being clustered are small sub-documents. The relevant portions of a relevant document might, indeed, form clusters with other similar sub-documents from the same or other relevant documents. But the non-relevant

portions of such documents may also form clusters. In such cases, we have no means of distinguishing the clusters; we must count a cluster as having a relevant "document" if any sub-document in the cluster came from a relevant document. This is clearly a flaw in how we measure relevance and it makes it difficult to assess the effectiveness of our processes.

4.3. Assessors' Choices

With respect to sub-document clusters, given the clusters we actually presented to assessors, did assessors select the "good" clusters? Were the presentations (in title-and-terms form or summary form) effective?

Our basic answer is *No*.

Tables 4a, 4b, and 4c, show in yellow the clusters that assessors chose as "On Topic." The values in the cells give the relevance density of the clusters; bold numbers indicate "good" density. In general, there is no indication that assessors could distinguish good groups from bad. Among quintads, assessors chose thirty-six (of sixty-eight available) good groups and thirty-seven bad ones. Among sub-document clusters (with seventy-eight available good groups), assessors chose twenty-five good groups and seventy-two bad ones using the titles & terms forms and twenty-two good and seventy-three bad groups using the summary forms. It is clear that the relevance density values do not pattern in any meaningful way with the clusters that the assessor chose.

Ranking of SUB-DOCUMENT CLUSTERS as Presented to Assessors

	421	430	448	425	427	415	409	422	428	424	416	408	410	429	432	406	404	407	418	417	402	446	447	411	431
1	1.00	0.30	0.10	0.70	0.10	0.71		0.80	1.00	0.20	0.50														
2	1.00	0.67	0.70	0.70	0.10	0.10	0.60			0.33		0.60	0.22	0.57	0.10			0.50							
3	0.40	1.00	0.50	0.10		0.60	0.90	0.10									0.30		0.70						
4	0.44	0.80	0.22	0.10			0.40						0.30			0.89						0.13			
5	0.43	0.90	0.70		0.10			0.80									0.20		0.10		0.11				
6	0.83	0.90	0.29		0.10		0.30		0.13		0.13	0.40		0.67	0.20					0.20					
7	0.20	0.20		0.60			0.90	0.10	0.70			0.20					0.20								
8	0.20	0.50	0.60		0.90	0.20	0.40	0.20	0.90	0.20	0.20		0.13	0.50	0.30	0.17		0.10							
9	0.10	0.30	0.30	0.50	0.40	0.40	0.80	0.40	1.00											0.20	0.10				
10	0.10	0.80			0.90	0.50	0.50	0.50	0.89			0.10	0.80										0.30		
11	0.89	0.20	0.11	0.40	0.60		0.20		0.50	0.10		0.10	0.20			1.00				0.20					
12		0.33	0.30			0.70	0.70	0.30	0.25		0.70		0.11												
13	0.80	0.70		0.10	0.30	0.88	0.60	0.11								0.22									
14	0.20	0.10		1.00	0.80		0.60	0.80		0.33		0.10	0.10		0.10										
15					0.50	0.20	0.80	0.50								0.43	1.00				0.10				
16	0.70		0.10		0.30	1.00		0.40		0.10		0.30	0.33	0.10						0.40					
17	0.70				0.10	0.40			0.30			0.30		0.11											
18	0.40		0.50	0.20		0.20	0.30		0.11	0.40			0.25	0.10		0.70								0.13	
19	0.40	0.10				0.40	1.00		0.25																
20	0.80		0.57			0.10	0.90		0.29		0.10	0.14		0.50		0.60	0.70								
21	0.70			0.10		0.20										0.90									
22			0.13	0.10		0.38	0.80		0.20			0.10				0.14			0.10	0.20					0.20
23	1.00	0.10	0.20		0.30	0.30	0.13		0.57			0.25													
24		0.20	0.20	0.10	0.80		0.90	1.00		0.11				0.20	0.20										
25	0.40		0.20	0.10	0.70	0.20	0.90					0.56	0.10		0.20										
26	0.40			0.10	0.10		0.30						0.13	0.11											
27	0.30		0.14		0.40			0.50			0.40			0.70	0.30	0.10									
28	1.00		0.30	0.50	1.00		0.10	0.90	0.40			0.22		0.10						0.50					
29	0.25		0.60	0.40	0.20		0.10	0.80	0.33					0.10	0.10										
30		0.10	0.50			0.80	0.20	0.20	0.20		0.10	0.20		0.17		1.00									
MRR	3.72	3.42	3.34	2.92	2.90	2.89	2.59	2.56	2.23	1.95	1.50	1.44	1.42	1.24	1.02	0.88	0.79	0.63	0.58	0.51	0.38	0.25	0.10	0.06	0.05

Cluster Density >= 0.8
0.8 > Cluster Density >= 0.4
0.4 > Cluster Density > 0
Clusters Visible to the Assessor

Table 3. Ranking of Sub-Document Clusters as Presented to Assessors

QUINTAD FORM

	421	430	427	409	428	448	415	408	416	432	410	429	422	407	418	425	446	447	406	404	431	417	402	424	411
1	0.40	0.60	0.80	0.20	0.80	0.40	0.40	0.60	0.20	0.20	0.60	0.80		0.40	0.40		0.20	0.20							
2	1.00	1.00	0.60	0.60	1.00	0.20		0.40												0.60	0.20				
3	0.40	1.00	0.60	0.80	0.60	0.20			0.40				0.40			0.20						0.60	0.20		
4	0.40	0.60	1.00	0.60			0.20					0.40	0.40			0.20			0.80						
5	0.40	0.80	0.60		1.00	0.20	0.80	0.60					0.20			0.80			0.20			0.40		0.20	
6	1.00	1.00	0.40	0.40	0.60	0.60				0.20	0.60					0.20									
7	0.80	0.40	0.20				0.40		1.00	0.20			1.00							0.20	0.20				
8	0.20	0.40	0.40	1.00							0.20	0.60				0.40			0.60						
9	0.60	0.20			1.00	0.40	0.40						0.20	0.20					0.40				0.20		
10	0.80	0.60	0.80	0.20						0.40	0.20		0.20		0.20										

Table 4a. Assessor-Selected Quintads

TITLES & TERMS FORM

	421	430	448	425	427	415	409	422	428	424	416	408	410	429	432	406	404	407	418	417	402	446	447	411	431
1	1.00	0.80	0.10	0.70	0.10	0.71		0.80	1.00	0.20	0.50														
2	1.00	0.67	0.70	0.70	0.10	0.10	0.60			0.33		0.60	0.22	0.57	0.10			0.50							
3	0.40	1.00	0.50	0.10		0.60	0.90	0.10									0.30		0.70						
4	0.44	0.80	0.22	0.10			0.40						0.30			0.89						0.13			
5	0.43	0.90	0.70		0.10			0.80									0.20		0.10		0.11				
6	0.83	0.90	0.29		0.10		0.30		0.13		0.13	0.40		0.67	0.20					0.20					
7	0.20	0.20		0.60			0.90	0.10	0.70			0.20					0.20								
8	0.20	0.50	0.60		0.90	0.20	0.40	0.20	0.90	0.20	0.20		0.13	0.50	0.30	0.17		0.10							
9	0.10	0.30	0.30	0.50	0.40	0.40	0.80	0.40	1.00											0.20	0.10				
10	0.10	0.80			0.90	0.50	0.50	0.50	0.89			0.10	0.60										0.30		
11	0.89	0.20	0.11	0.40	0.60		0.20		0.50	0.10		0.10	0.20			1.00				0.20					
12		0.33	0.30			0.70	0.70	0.30	0.25		0.70		0.11												
13	0.80	0.70		0.10	0.30	0.88	0.60	0.11								0.22									
14	0.20	0.10		1.00	0.80		0.60	0.80		0.33		0.10	0.10		0.10										
15					0.50	0.20	0.80	0.50								0.43	1.00				0.10				

Table 4b. Assessor-Selected Clusters on Titles & Terms Forms

SUMMARY FORM

	421	430	448	425	427	415	409	422	428	424	416	408	410	429	432	406	404	407	418	417	402	446	447	411	431
1	1.00	0.30	0.10	0.70	0.10	0.71		0.80	1.00	0.20	0.50														
2	1.00	0.67	0.70	0.70	0.10	0.10	0.60			0.33		0.60	0.22	0.57	0.10			0.50							
3	0.40	1.00	0.50	0.10		0.60	0.90	0.10									0.30		0.70						
4	0.44	0.80	0.22	0.10			0.40						0.30			0.89						0.13			
5	0.43	0.90	0.70		0.10			0.80									0.20		0.10		0.11				
6	0.83	0.90	0.29		0.10		0.30		0.13		0.13	0.40		0.67	0.20					0.20					
7	0.20	0.20		0.60			0.90	0.10	0.70			0.20					0.20								
8	0.20	0.50	0.60		0.90	0.20	0.40	0.20	0.90	0.20	0.20		0.13	0.50	0.30	0.17		0.10							
9	0.10	0.30	0.30	0.50	0.40	0.40	0.80	0.40	1.00											0.20	0.10				
10	0.10	0.80			0.90	0.50	0.50	0.50	0.89			0.10	0.60										0.30		
11	0.89	0.20	0.11	0.40	0.60		0.20		0.50	0.10		0.10	0.20			1.00				0.20					
12		0.33	0.30			0.70	0.70	0.30	0.25		0.70		0.11												
13	0.80	0.70		0.10	0.30	0.88	0.60	0.11								0.22									
14	0.20	0.10		1.00	0.80		0.60	0.80		0.33		0.10	0.10		0.10										
15					0.50	0.20	0.80	0.50								0.43	1.00				0.10				

Table 4c. Assessor-Selected Clusters on Summary Forms

5. Conclusions

In general, our TREC-2004 Hard-Track experiments demonstrated the basic utility of our baseline and quintad processing for information retrieval tasks. However, our attempts to improve the quality of clustering (and concentration of relevant information) through tighter sub-document clustering were not successful.

Both quintad processing and clustering on sub-documents did succeed in grouping relatively more relevant documents in a relatively small numbers of groups. Furthermore, our ranking of the clustered groups (to present to assessors) did largely succeed in making some—frequently many—"good" groups available for judgment and feedback. Nevertheless, neither the titles & terms versions of our groups (both quintads and sub-document clusters) nor the summary versions (of sub-document clusters) seems to have provided assessors with the information required to distinguish clusters with relevant documents from those with no such documents.

We attribute the stable (stronger) performance of the quintad approach to the fact that we did re-rank clusters (and their contained documents) such that more relevant documents (among the top-200) were concentrated near the top (visible) portion of the response set. In short, almost any assessor choices among the reduced set of documents had a reasonable chance of resulting in some relevant documents for feedback. In the case of (small) sub-document clusters, however, we generated many small passages and the number of such clusters and their contained sub-documents varied widely. We believe we have not yet identified the appropriate strategy for grouping small sub-documents, scoring (ranking) such clusters, and representing them to assessors. Such work will be the focus of our future efforts.

References

[Evans & Lefferts 1994] Evans, D.A.; Lefferts, R.G**.** Design and Evaluation of the CLARIT–TREC-2 System. In Donna Harman (Editor), *The Second Text REtrieval Conference (TREC-2*). NIST Special Publication 500-215. Washington, DC: U.S. Government Printing Office, 1994, 137–150.

[Evans & Lefferts 1995] Evans, D.A.; Lefferts, R.G. CLARIT–TREC Experiments. *Information Processing and Management*, Vol.31, No.3, 385–395, 1995.

[Milic-Frayling et al. 1997] Milic-Frayling, N.; Zhai, C.; Tong, X.; Jansen, P.; Evans, D.A. Experiments in Query Optimization. The CLARIT System TREC-6 Report. In E.M. Voorhees and D.K. Harman (Editors), *The Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication 500-240. Washington, DC: U.S. Government Printing Office, 1998, 415–454.

[Shanahan et al. 2004] Shanahan, J.G.; Bennett, J.K.; Evans, D.A.; Hull, D.A.; Montgomery, J. Clairvoyance Corporation Experiments in the TREC 2003 High Accuracy Retrieval from Documents (HARD) Track. In E.M. Voorhees and Lori P. Buckland (Editors), *The Twelfth Text Retrieval Conference, TREC-2003.* NIST Special Publication 500-255. Washington, DC: U.S. Government Printing Office, 2004, 152–160.