Clairvoyance Corporation Experiments in the TREC 2003 High Accuracy Retrieval from Documents (HARD) Track

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

Clairvoyance Corporation, Pittsburgh, PA
{jimi, jfb, dae, hull, jmontgomery}@clairvoyancecorp.com

1. Introduction

The Clairvoyance team participated in the HARD Track, submitting fifteen runs. Our experiments focused primarily on exploiting user feedback through clarification forms for query expansion. We made limited use of the genre and related text metadata. Within the clarification form feedback framework, we explored the cluster hypothesis in the context of relevance feedback. The cluster hypothesis states that closely associated documents tend to be relevant to the same requests [Van Rijsbergen, 1979]. With this in mind we investigated the impact on performance of exploiting user feedback on groups of documents (i.e., organizing the top retrieved documents for a query into intuitive groups through agglomerative clustering or document-centric clustering), as an alternative to a ranked list of titles. This forms the basis for a new blind feedback mechanism (used to expand queries) based upon clusters of documents, as an alternative to blind feedback based upon taking the top *N* ranked documents, an approach that is commonly used.

Though our submitted results suffered from incorrect reference statistics in the baseline run the cluster hypothesis was validated; feedback through our cluster-based clarification form yielded a 20% improvement for mean average precision over blind feedback. The cluster hypothesis was further validated, in a somewhat ideal setting, when expansion was performed automatically using the optimal cluster that was selected using a post-hoc analysis. Here, the boost in performance over blind feedback is 20% and is comparable to the TREC max for this track. Ongoing work is investigating techniques that would automatically select the optimal cluster(s).

2. Approach to High Accuracy Retrieval from Documents (HARD) Task

2.1 HARD Corpus

The HARD evaluation corpus is composed of documents from The New York Times (NYT), Associated Press Worldstream (APW), Xinghua English (XIE),The Congressional Record (CR), and Federal Register (FR). We merged all corpora to form one large corpus over which global statistics, such as inverse document frequency (IDF), were computed. The CLARIT system is based on passage retrieval. Passages are defined at indexing time and are commonly referred to as sub-documents (or SubDocs). The sub-document size is fully configurable, with the default setting (used here) producing passages that range in size from 8 to 20 sentences. The typical default sub-document size is 12 sentences. Using this process, we split the documents in this evaluation corpus into passage-size sub-documents.

2.2 Query Formulation and Blind Feedback

A HARD topic is composed of a *title* field, a *description* field and other metadata. From the other meta-data, we considered only *genre* and *related text* in our experiments. To form a query for a topic we merged the title and description fields. We extracted terms from the query text using the Clarit system natural language processing, giving morphologically-normalized words, phrases, and sub-phrases as index terms [Evans and Lefferts, 1995]. Each query term, *t*, is associated with a weight calculated as follows:

$$Weight(t) = TF(t) * IDF(t) * coefficient(t)$$

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

$$TF(t) = 0.5 + 0.5 * TermFreq(t)$$

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:

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

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*.

We submit this query to our Clarit retrieval engine and get a ranked list of sub-documents based upon the score between each sub-document and the query. This sub-document score is calculated as follows:

$$Score(SDoc, Query) = \sum_{t \in Query} TF_{SDoc}(t) * IDF(t)^2 * coefficient(t) * TF_{Query}(t)$$

This ranked list of sub-documents is post processed such that sub-documents belonging to a single document are reduced to the original document and its score is set to the score of the highest scoring sub-document.

Subsequently, blind feedback can be used to expand the original query automatically using terms extracted from the top *C* ranked documents. Blind feedback has been shown to improved ah-hoc retrieval performance [Evans and Lefferts, 1995]. A similar process can be used for supervised/directed feedback. Terms are extracted from all sub-documents that score at or above the *C*-th document score. This may lead to using multiple sub-documents from the same document. Extraction of terms from these top sub-documents is performed using Clarit NLP. Term selection is performed using the following steps. Terms are ranked in decreasing order using the *Prob2* weighting scheme, which is defined as follows:

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

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 appended to the original query. We set term coefficients in the expanded query as follows:

- terms that occur in both the expanded set and the original query are set to 1.5;
- terms that occur in the query only are set to 1.0; and
- terms that occur in the expanded set only are set to 0.5.

In all our experiments, unless otherwise noted, we set *C* to 6 (documents) and *k* to 30 (terms).

In our post-TREC experiments, we set the coefficients of query terms and new terms through a term normalization algorithm:

- the coefficient of terms that occur in the expanded set and in the original query are set to *(1 * boostFactor) + normalisedProb2*;
- terms that occur in the query only are set to 1.0; and
- terms that occur in the expanded set are set to *normalisedProb2*.

For our experiments, *boostFactor* was set to 2. The *normalisedProb2* factor is calculated as follows:

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

2.3 Clarification Forms

We explored two types of clarification forms. Both forms presented documents in groups derived from clustering. The first form, called as

the *title-based form*, presents the top ranking documents for a query in groups. These groups are formed by clustering the top 100 ranked results using a simple clustering strategy outlined below. Each group is presented using a list of terms, corresponding to typical terms for the group, and a list of documents, where each document is represented using its title and the information source (very similar to the forms used in Scatter-Gather [Hearst and Pederson, 1995]). A clarification form of this type corresponding to Topic 33 is depicted in Figure 1.

The user is presented with ten groups of documents for each topic, where each group consists of five documents: the seed document and its four nearest neighbors. The seed for the first group is set to the top ranking document. This seed is then used to rank the top 100 documents. The top 4 ranking documents, in addition to the seed document, comprise the first group. Other groups are formed in a similar way, where the seed is set to be the top-most ranking document that has not already been used as a seed or as a nearest neighbor.

The user is asked to judge the relevance of each group for a query as being "*On Topic*", "*Not on Topic*", "*Unsure*" 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 user believes none of the group documents or group terms represent the topic well. If uncertain whether the group accurately represents the topic, the user could choose "*Unsure*." And if the user runs out of time or simply fails to judge a group, the default value for a group is "*Unjudged*".

The system uses the group judgments, as directed feedback, to expand the query with terms extracted from the constituent documents of the positively assessed groups. We add all documents from groups that are marked "*On Topic*" to the feedback pool. If the number of groups marked as "*On Topic*" is less than two, we also include the "*Unsure*" groups in the feedback pool. If there are still fewer than two groups, we use the default blind feedback strategy (of 6 documents and 30 terms). If there are enough documents in the feedback pool, we expand the original query using the expansion strategy defined above, where *C* is set to the number of documents in the pool.

We subsequently perform a retrieval over the entire corpus using the expanded query. We post-process this ranked list by front-loading (promoting to the top of the list) all documents in the feedback pool, and excluding documents from groups marked as "*Not On Topic*".

In the second form documents are again organized into groups. Here, each group is represented using a list of terms, corresponding to typical terms from the documents that make up the group. The groups in this form are created by clustering the top-ranking 100 documents for the original query (using a version of the agglomerative average-link clustering algorithm). The terms are extracted from the documents in each cluster using Clarit NLP and the top forty terms (determined by their Prob2 weight) are listed as representative terms for the cluster.

Once again, the user is asked to judge the relevance of each group for the query as being "*Clearly Related*", "*Somewhat Related*", "*Not Related*", "*Unsure*", or "*Unjudged*". The "*Clearly Related*" option denotes that some of the terms in the group were strongly representative of the topic. The "*Somewhat Related*" option signifies that some of the terms in the group were somewhat representative of the topic. The "*Not on Topic*" option indicates that none of the terms was on topic. The other options are self-explanatory.

The system uses the group judgments as a from of directed feedback to expand the query with terms extracted from the constituent documents of the positively assessed groups. We add all documents from groups that are marked "*Clearly Related*" to the feedback pool. If the number of groups marked as "*Clearly Related*" is less than two, we also include the "*Somewhat Related*" groups in the feedback pool. If there are still fewer than two groups, we use the default blind feedback strategy to expand the query. If we have enough documents in the feedback pool, we expand the original query using the expansion strategy defined above, where *C* is set to the number of documents in the pool.

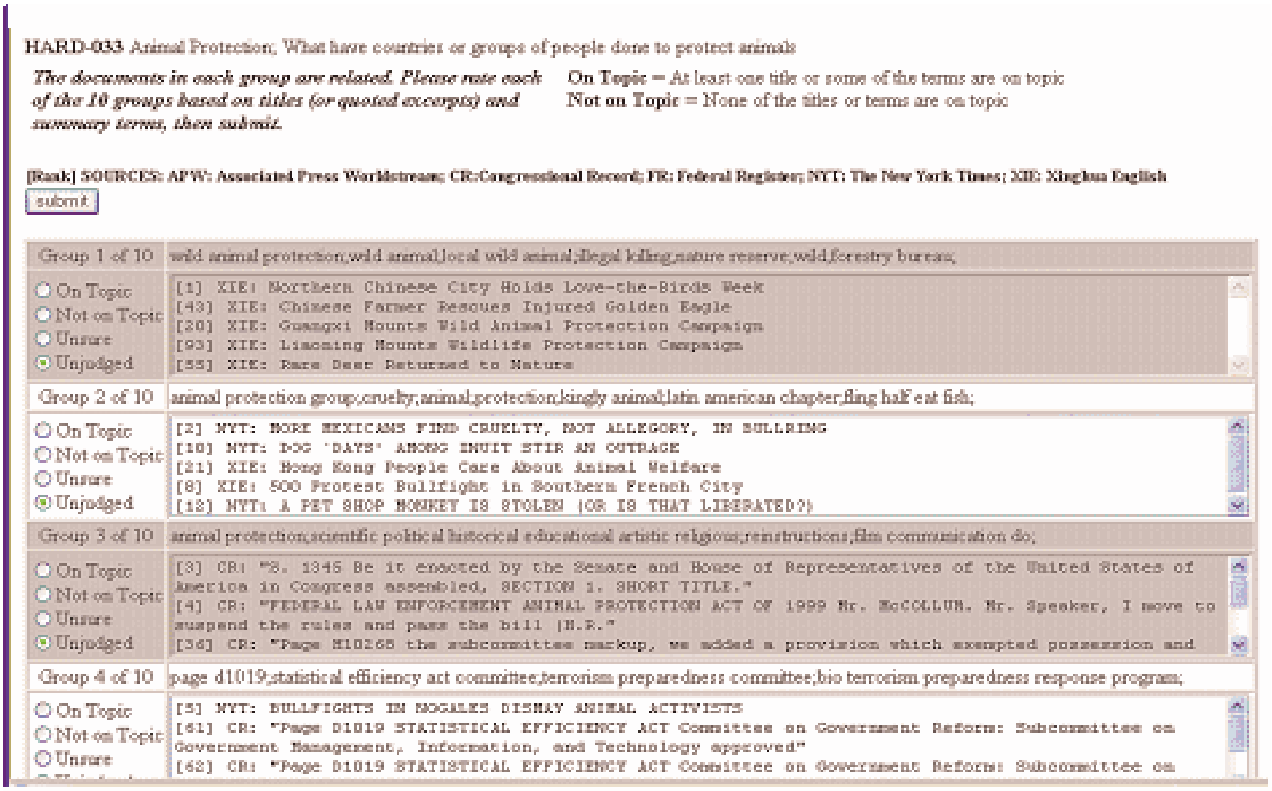

Figure 1. An example of a title-based clarification form using typical terms and document titles.

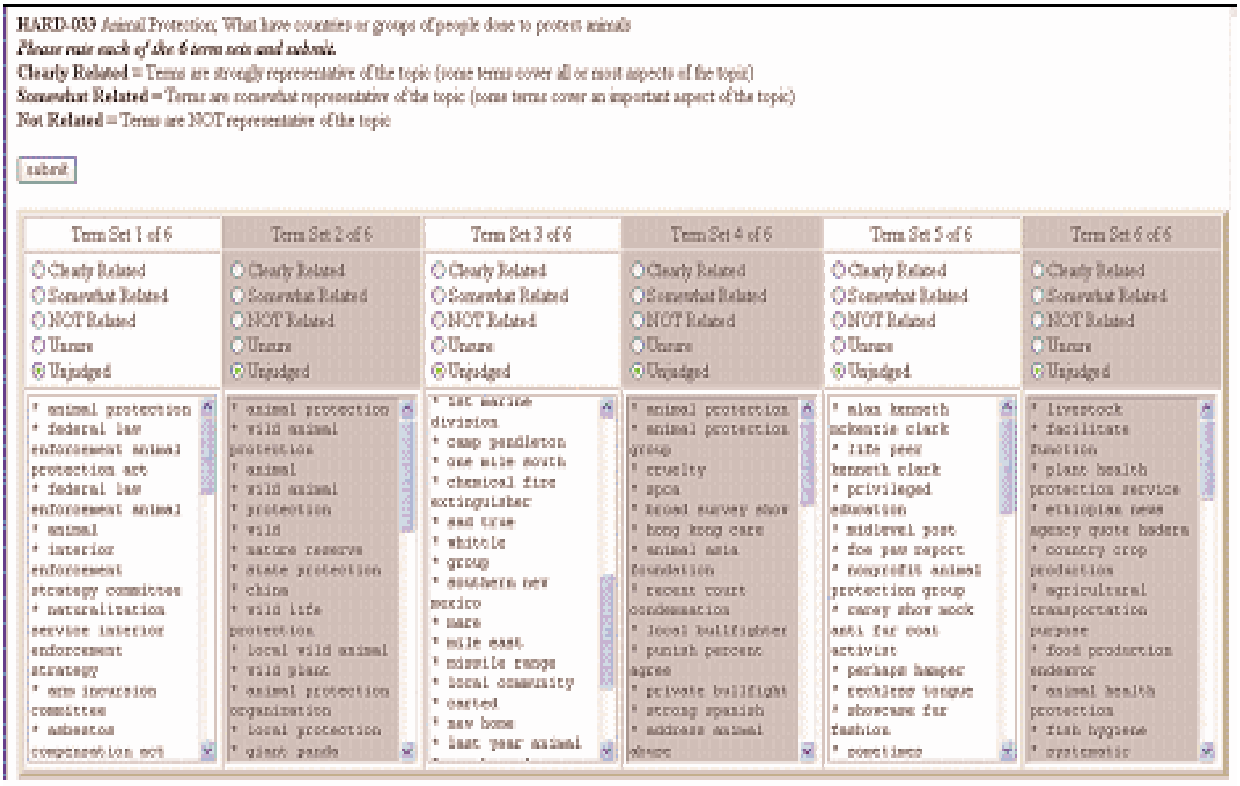

Figure 2. An example of a term based clarification form.

We subsequently perform a retrieval over the entire corpus using the expanded query. We post-process this ranked list by upgrading all documents that occur in groups marked as "*Clearly Related*" to the front of the ranked list, while documents that are members of groups marked as "*Not Related*" are excluded from the top 1000 documents list.

2.4 Exploiting Topic Metadata

We selectively used two pieces of topic metadata: genre and related text. The genre metadata field introduces a strong source-specific bias for individual topics. We factor in the topic genre with a post-processing step to re-rank the retrieval results. Essentially, we assign a weight factor ranging from 0 to 1 for each genre/source pair. The weight factor is a rough measure of the relative likelihood that a document from that source would be relevant to the given topic genre. This weight is multiplied by the final retrieval score and the documents are re-ranked accordingly. The final weight table is given below:

Genre\Source	APW	CR	FR	NYT	XIE
Any	1.0	1.0	1.0	1.0	1.0
Administrative	0.2	1.0	1.0	0.2	0.2
I-Reaction	0.5	0.2	0.2	0.5	1.0
Overview	1.0	0.2	0.2	1.0	1.0
Reaction	1.0	0.2	0.2	1.0	1.0

It might have been possible to learn these weights automatically given training data, but we decided that there were not enough training data available to make this a productive exercise.

We used the related text to expand the original query (i.e., we concatenated it to the title and description data). For some experiments, where the related text contained URLs, we fetched the content of the page pointed to by the URL and concatenated that to the original query also.

3. Results

Table 1 presents a list of the experiments we carried out. Each row in this table denotes an experiment and how the original query was constructed (using related text or downloaded related web pages that were pointed to in the related text data), and whether any genre post-processing was done. The experiments labeled *NewBase630*, *BestClusterRun*, and *BestGroupRun* correspond to experiments that were performed after the submission deadline.

Table 2 presents the results of our experiments in terms of mean average precision, exact precision, and precision at 10.

Overall, our submitted runs suffered from incorrect reference statistics in the baseline run, which was also used to generate both clarification forms. The mean average precision (MAP) over the top 1000 documents for this baseline submission was 0.24, denoted as experiment *CLSTD630* in Table 2. This corresponds to the median results for all systems in this track. This baseline run has since been improved to 0.31 (experiment *NewBase630*). Figure 3 presents a comparison of TREC Median and Maximum to our submitted baseline run and our post-TREC baseline run (*NewBase630*).

When we incorporated feedback from the title-based form, the MAP improved to 0.28 from 0.24. The term-based form did not yield any significant improvement. Figure 4 presents a comparison of the TREC Median and Maximum to our submitted baseline run and our best submitted run (*CLAI1NG*).

Our approaches to exploiting the topic metadata (genre and related text) do not yield any improvement. Table 3 presents passage level evaluation results. Our passage level results reflect our document level results in terms of the evaluation measures.

We performed various follow-up experiments where we exploited the cluster hypothesis. Here, our baseline performance corresponded to choosing a single best-performing group/cluster for each topic automatically. When documents were grouped using agglomerative clustering, and the best performing cluster was selected (post-hoc), we attain an overall MAP of 0.37 (*BestClusterRun*). Similarly, when grouping using single-ranked documents (as was used in generating the groups in the title-based form), using the best performing cluster selected, we attain an overall MAP of 0.37 (*BestGroupRun*). These experiments compare very favorably to the TREC maximum MAP of 0.40 for this track.

Submission name	Genre (MetaData)	Related Text (MetaData)	Related Text (Web Pages)
CLAI1NG	No	No	No
CLAI1G	Yes	No	No
CLAI2NG	No	No	No
CLAI2G	Yes	No	No
CLAI2RTNG	No	Yes	No
CLAI2RTG	Yes	Yes	No
CLAI2WRTNG	No	Yes	Yes
CLAI2WRTG	Yes	Yes	Yes
CLAISTDNG	No	No	No
CLAISTDG	Yes	No	No
CLAISTDRTG	Yes	Yes	No
CLAISTDRTNG	No	Yes	No
CLAISTDWRTG	Yes	Yes	Yes
CLAISTDWRTNG	No	Yes	Yes
CLSTD630	No	No	No
NewBase630	No	No	No
BestClusterRun	No	No	No
BestGroupRun	No	No	No

Table 1. Submission and post-submission experiment details. (All runs used Title+Description fields as the query.)

Experiment	Avg Prec	R-Prec	Prec @ 10
CLAI1G	0.2884	0.3229	0.4729
CLAI1NG	0.2917	0.3246	0.4729
CLAI2G	0.2499	0.2861	0.4188
CLAI2NG	0.2514	0.2870	0.4146
CLAI2RTG	0.2218	0.2634	0.4125
CLAI2RTNG	0.2225	0.2633	0.4083
CLAI2WRTG	0.2138	0.2533	0.4104
CLAI2WRTNG	0.2170	0.2569	0.4063
CLAISTDG	0.2309	0.2658	0.3750
CLAISTDNG	0.2345	0.2712	0.3917
CLAISTDRTG	0.2334	0.2686	0.3979
CLAISTDRTNG	0.2285	0.2618	0.4000
CLAISTDWRTG	0.2134	0.2468	0.3729
CLAISTDWRTNG	0.2105	0.2420	0.3646
CLSTD630	0.2341	0.2772	0.3938
NewBase630 (post-TREC run)	0.3069	n/a	n/a
BestClusterRun (post-TREC run)	0.3727	n/a	n/a
BestGroupRun (post-TREC run)	0.3741	n/a	n/a
TREC median	0.2841	0.2994	0.4729
TREC max	0.4069	0.4250	0.6500

Table 2: Document Level Evaluation Results.

Run	F @ 30	R-Prec	Prec @ 10
CLAI1G	0.0905	0.2426	0.3235
CLAI1NG	0.0851	0.2266	0.2886
CLAI2G	0.0814	0.1900	0.2538
CLAI2NG	0.0781	0.1815	0.2201
CLAI2RTG	0.0762	0.1773	0.2507
CLAI2RTNG	0.0728	0.1708	0.2176
CLAI2WRTG	0.0761	0.1733	0.2483
CLAI2WRTNG	0.0728	0.1704	0.2142
CLAISTDG	0.0738	0.1799	0.2230
CLAISTDNG	0.0733	0.1737	0.2056
CLAISTDRTG	0.0826	0.2076	0.2610
CLAISTDRTNG	0.0745	0.1782	0.2246
CLAISTDWRTG	0.0822	0.1963	0.2582
CLAISTDWRTNG	0.0744	0.1700	0.2243
CLSTD630	0.0700	0.1726	0.2100
BestClusterRun (post-TREC run)	n/a	n/a	n/a
BestGroupRun (post-TREC run)	n/a	n/a	n/a
TREC med	0.1000	0.1794	0.2574
TREC max	0.1738	0.3195	0.3973

Table 3: Passage level evaluation results.

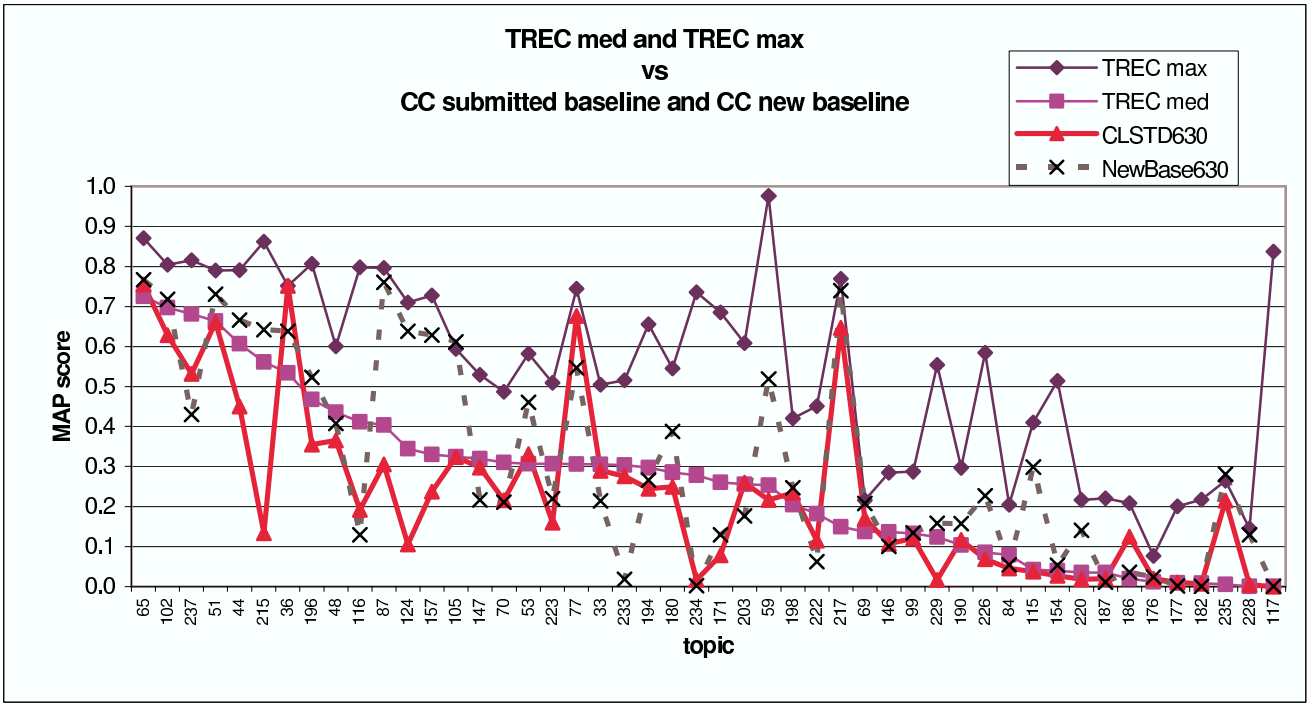

Figure 3: Comparison of TREC Median and Maximum to Clairvoyance's submitted baseline run and Clairvoyance's post-TREC baseline run.

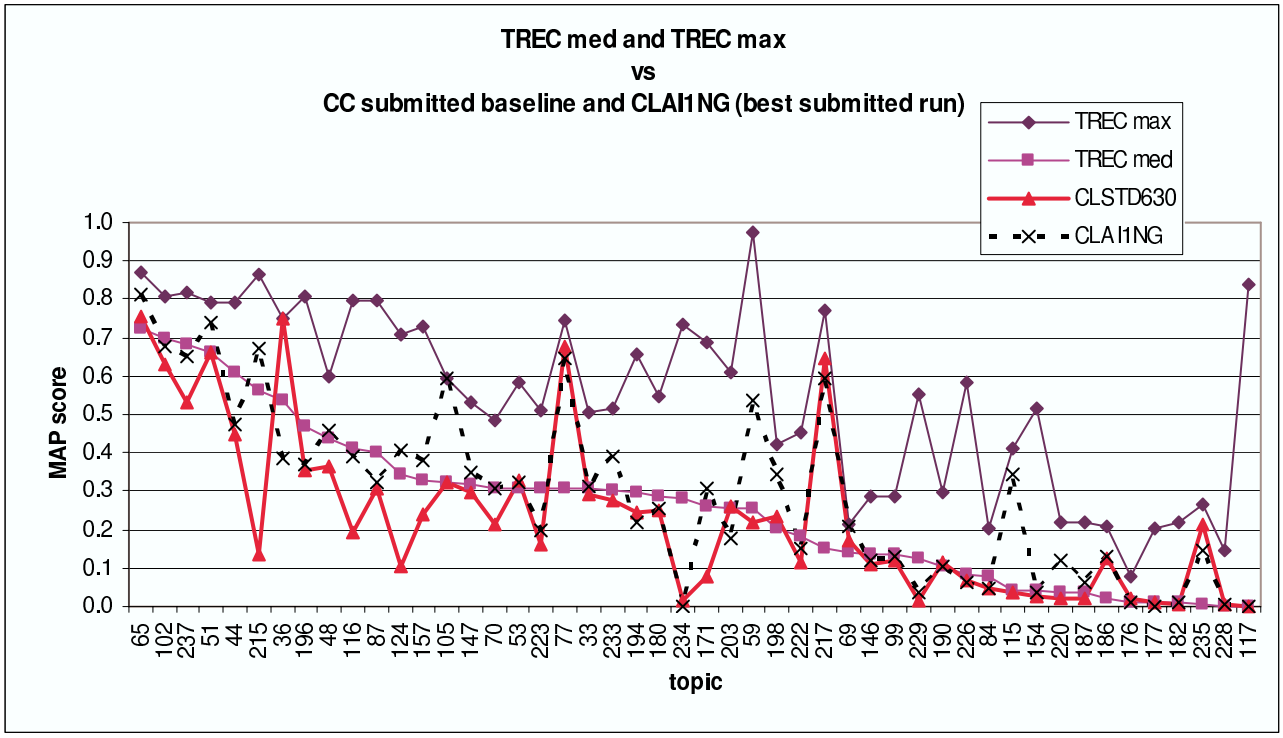

Figure 4: Comparison of TREC Median and Maximum to Clairvoyance's submitted baseline run and Clairvoyance's best submitted run (CLAI1NG).

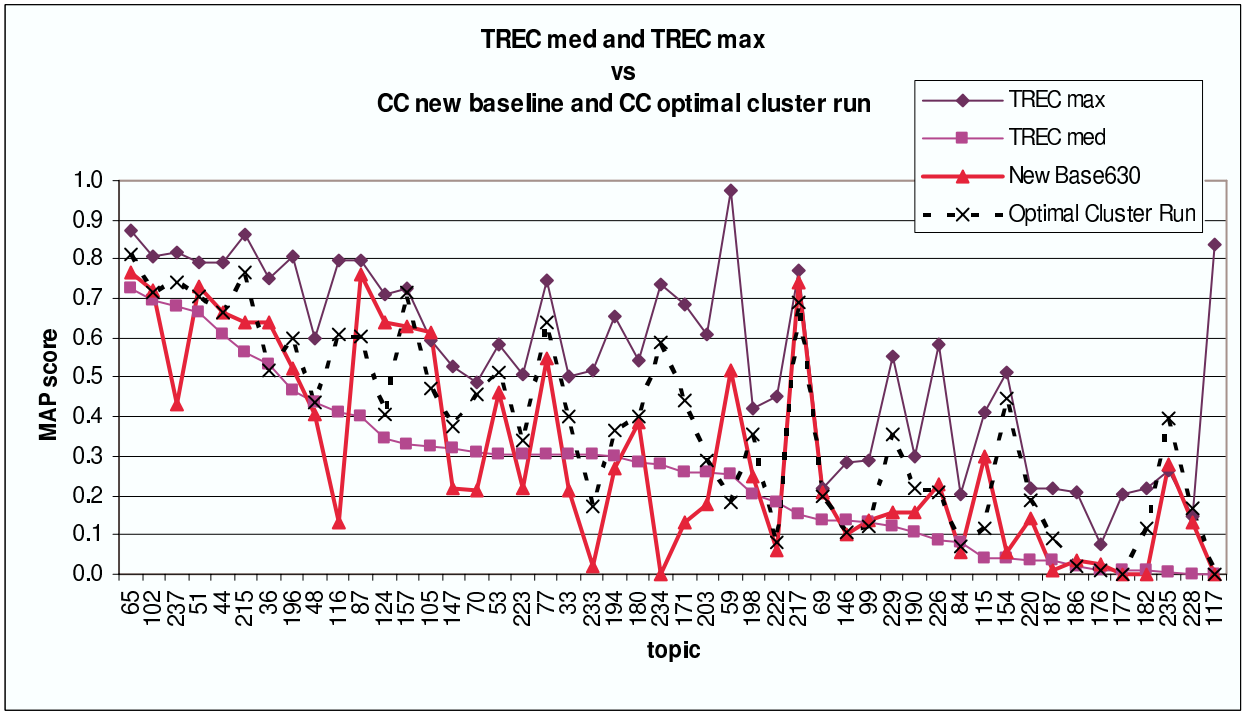

Figure 5: Comparison of TREC Median and Maximum to Clairvoyance's Post-TREC Baseline Run and Clairvoyance's Post-TREC Optimal Cluster Run.

Though we have investigated a number of approaches to automatically selecting a group(s), none of our examined approaches seems to provide consistent results across all topics.

Figure 5 presents a comparison of the TREC Median and Maximum to our post-TREC baseline run and our post-TREC optimal cluster run. It is important to note that this 0.37 MAP performance comes from selecting the best group. Performance can be potentially improved by incorporating more than one group as feedback. We are currently exploring other approached on how to select these groups for feedback automatically.

3. Conclusions

For our HARD experiments we explored the cluster hypothesis in the context of manual feedback through clarification forms and automatic feedback. We compared both of these approaches to blind feedback.

Though our submitted results suffered from incorrect reference statistics in the baseline run, the cluster hypothesis was validated; feedback through our cluster-based (title) form yielded a 20% improvement for mean average precision over blind feedback. While we have demonstrated the benefits of manually selecting the optimal clusters for feedback, a better comparison would be to have the user manually select documents from a ranked list of titles. This could potentially provide comparable results without the burden of clustering.

The cluster hypothesis was also validated, in a somewhat ideal setting, when expansion was performed automatically using the optimal group that was selected using a post-hoc analysis. Here, the boost in performance over blind feedback is 20% and is comparable to the TREC max for this track. Our continuing work is investigating techniques that would automatically select the optimal cluster(s).

References

[1] Evans, David.A., and Robert G. Lefferts. CLARIT–TREC Experiments. Information Processing and Management, Vol.31, No.3, pp.385–395, 1995.

[2] Hearst, Marti A., and Jan O. Pedersen. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval.

[3] Van Rijsbergen, C. J. Information Retrieval, Butterworths, London, Second Edition, 1979.