# Relevance Feedback Track Overview: TREC 2008

Chris Buckley, Sabir Research
Stephen Robertson, Microsoft

## 1 Introduction

Relevance Feedback has been one of the successes of information retrieval research for the past 30 years. It has been proven to be worthwhile in a wide variety of settings, both when actual user feedback is available, and when the user feedback is implicit.

However, while the applications of relevance feedback and type of user input to relevance feedback have changed over the years, the actual algorithms have not changed much. Most algorithms are either pure statistical word based (for example, Rocchio or Language Modeling), or are domain dependent. We should be able to do better now, but there have been surprisingly few advances in the area.

In part, that's because relevance feedback is hard to study, evaluate, and compare. It is difficult to separate out the effects of an initial retrieval run, the decision procedure to determine what documents will be looked at, the user dependent relevance judgment procedure (including interface), and the actual relevance feedback reformulation algorithm. Setting up a framework to look at these separate effects for future research is an important goal for this track.

Why now? We have a lot more natural language tools than we had 10 or 20 years ago. We're hopeful we can get people to actually use those tools to suggest what makes a document relevant or non-relevant to a particular topic.

The question-answering community has been very successful at categorizing questions and taking different approaches for different categories. The success has not transferred over to the IR task, partly because there simply isn't enough syntactic information in a typical IR topic to offer clues as to what is wanted. But given relevant and non-relevant judgments, it should be much easier to form categories for topics (e.g., this topic requires these two aspects to both be present, while this other topic does not), and take different approaches depending on topic.

Relevance feedback is an area that's ripe for major advances, and is being held back because there isn't a common methodology for investigating and comparing relevance feedback algorithms. This track establishes that methodology, and offers groups the ability to evaluate and compare their relevance feedback algorithms.

# 2 Goals and procedures for Relevance Feedback 2008

There were 3 main goals for this track:

1. Target evaluating and comparing just the RF algorithm - all groups will work with exactly the same relevance judgments. Hopefully compare both statistical and NLP intensive use of relevance information (what makes a document relevant).

2. Establish good baseline RF results for multiple amounts of relevance info. There will be separate runs which contain (for each topic):

   - A. no relevance info (baseline retrieval)
   - B. 1 relevant document
   - C. 3 relevant documents and 3 non-relevant documents
   - D. 10 judged docs (superset of C, so at least 3 relevant and 3 non-relevant documents )
   - E. large amounts of judged documents (40 to 800)

3. Try to establish, for these runs, the amount of improvement possible with more relevance info.

## 2.1 Collection

The experimental test bed was the Terabyte doc collection, with 50 topics drawn from 2004-2006 Terabyte track, and 214 topics 2007 Million Query (MQ) track.

All topics were even-numbered topics from those tracks. Participating groups were allowed to develop and train on the odd-numbered topics, but not the even-numbered topics, to avoid over-training on the test sets.

There were 264 total topics. All were reasonably short (most between 1 and 8 words) natural language statements - the topics drawn from the Terabyte tracks were title-only topics. They include

1. 50 topics from the Terabyte tracks, for which exists several hundred judged docs available for each topic

2. 214 topics from the 2007 MQ track, for which exists 40-100 judged documents per topic.

There were 25 of these topics designated as a manual topic subset, 13 from the Terabyte topics and 12 from the MQ topics. Participating groups, if they wanted to, could have a human look at both those topics and initial results. Only one group did so (AmsterdamR), so the manual runs will not be further considered here.

### 2.2 Input judged documents

The judged docs used as RF input in Sets B, C, and D in goal 2 above, are the docs with the highest median retrieval ranks in the appropriate old track (either Terabyte 06 or MQ 07), using only the best run per group in the old track. Set C is a superset of Set B, Set D is a superset of C, and Set E is a superset of D.

The judged docs were distributed as 4 separate TREC qrels file, one for each of Sets B, C, D, E. Each qrels file has judgments for all 264 topics, with relevance values of 0 indicating nonrelevant, 1 indicating relevant, and 2 indicating highly relevant.

### 2.3 Submitted runs

Participants submitted 1 run on all 264 topics from each of the 5 Sets A-E, thus a basic standard group submission was 5 runs. Each run submission was in standard ranked TREC results format, and contained up to the top 2500 retrieved docs. The main evaluation was on the top 1000 retrieved docs which were not part of the RF input for any set of judgments for this topic (NIST removed Set E documents from the submitted runs before evaluation - the individual groups did not remove Set E docs from their submitted runs).

In addition, if a group submitted a full set of runs on Sets A-E, they could also submit a second set of runs on Sets B-E in order to compare 2 variants of their RF algorithm.

To ensure that runs were properly linked with their RF input, they were required to be named as basename.XN where 'X' was one of A-F depending on the RF input, and 'N' was either 1 or 2 depending on whether this was the first or second set of runs for the group. Thus "SabRF08.C1" was the Sabir group's first run using qrels Set C as the known relevant documents.

## 3 Evaluation

RF evaluation is tough since it's a real world process that we need to break up into manageable chunks to fairly compare runs. One major issue is what to do with the documents that the user has already seen and judged - i.e., the input docs to the RF algorithm. There's many proposed solutions to this in the literature; the best one for evaluating just RF algorithms (as is being done this year), is residual collection evaluation: remove from the collection both the relevant and nonrelevant docs that are input to the RF algorithm. Then normal evaluation methodology can be used without worrying about the evaluation effect of those input judged docs.

That works well for comparing runs that use the same input docs. However, if you want to compare runs using different sets of input docs, then the same principle says you have to expand the removed docs to be the union of the different sets of input docs. Otherwise, you are comparing runs made on residual collections which may contain very different numbers of relevant docs for a topic; that's quite problematic.

For the RF track this year, one major goal is to compare the effects of using increasing amounts of relevance info: for instance a set B run will be compared against a set E run.

Using a residual collection approach, that means removing set E documents from both runs before comparing.

There is a danger in removing too many documents from the result set - it may strongly increase the measurement error on some topics. If there are very few relevant docs left to be retrieved after removal of possibly hundreds of docs, then no evaluation measure is going to accurately compare systems on those topics; the notion of relevance is not generally accurate enough to support such comparisons. That will be looked at later.

## 3.1 Pooling and official evaluations

There were 2 separate evaluations done. Both of them pooled and evaluated not only the 118 RF runs, but also the 25 MQ runs submitted in June (so we have outside base cases for comparison). All official evaluations on all runs had Set E docs removed before any pooling or evaluations took place.

The first evaluation was MQ-style, targeting up to either 32 or 64 (50-50 split) documents for up to 239 topics. There were two different MQ measures being calculated: statMAP from NEU (on 208 topics), and expectedMAP from UMass (on 237 topics). These measures are intended to give the same ranking as MAP would if the runs had been fully judged, but algorithmically sample only a small number of docs. The purpose of MQ style evaluation is to be able to evaluate a much larger number of topics for the same judging effort as the usual TREC topN pooling. That's important for RF, since topic variability of results is not only affected by the normal inherent topic difficulty, and user interpretation of relevance (both always present in ad hoc evaluations), but also whether the docs used as RF input are representative. The topics potentially to be evaluated in the MQ style were 214 topics originally retrieved in the TREC MQ 2007 track, plus 25 topics from the 3 years of TREC Terabyte tracks. (Two topics were not judged in the MQ evaluation.) On average, 39 documents were judged per topic.

The second evaluation was a normal TREC pooling evaluation, initially planned to be done on the other 25 topics from Terabyte track (all submission were of 264 topics). Due to a mistake on the track coordinator's part, the set of 25 Terabyte topics to be evaluated in the pooling evaluation ended up being the same initial set as was judged in the MQ-style evaluation. There was extra assessor time available after the two evaluations; that time was used to judge a few more topics - thus of the original 50 Terabyte topics for which groups submitted results, 25 were judged in both evaluations, 6 were judged only in the pooling evaluation (thus a total of 31 pooling topics), and 19 were not judged at all. Given the limited resources available, the pool of documents to be judged for each topic consisted of the top 10 documents from every run. With overlap, that amounted an average of 357 documents per topic being judged.

The Pool10 evaluation is an approximation of the normal TREC evaluation strategy, and should allow ranking of systems by any of the standard evaluation measures. As always, values of measures may be different than if full judging were done, but system comparisons should still be valid. This evaluation should allow investigation of whether the effects of the RF were concentrated on just the top retrieved docs, or were more recall oriented. (One ever

| Evaluation | Num Topics | Total Judged | Total Rel | Min Num Rel | Max Num Rel |
|---|---|---|---|---|---|
| Pool10 | 31 | 11058 | 1723 | 4 | 177 |
| MQ-style | 237 | 9312 | 2386 | 0 | 46 |

Table 1: Statistics on Relevance Feedback Evaluations

present RF question for a particular system is whether the benefit is due to just finding a couple of good query expansion terms, or due to a lot of expansion terms establishing a useful context). The MQ evaluation measures should not be used for this sort of investigation.

# 4 Meta-evaluation results

Table 1 gives overall statistics for the two/three poolings and evaluations. The MQ-style evaluation effectively includes two different poolings of its own, one for each of the measures.

Out of the 31 Pool10 topics, 25 had more than 20 relevant. In the MQ-style evaluation, 29 topics had no judged relevant documents.

## 4.1 Duplicate judgments

Among the 25 topics that were judged in both evaluations, there were 767 documents judged twice. Those documents occur in both the distributed prels file and the qrels file. The assessors both agreed that 85 documents were relevant and 567 were non-relevant. They disagreed on 115 documents. Thus of the 200 documents at least one assessor judged relevant, there was agreement on 42%. That's consistent with all previous studies on expected assessor agreement.

## 4.2 MQ runs and Relevance Feedback runs

Pools for the two evaluations included both the 118 Relevance Feedback Track runs and the 25 MQ Track runs (submitted in June). The purpose of including the MQ runs in the relevance feedback pools was three-fold. First, the June MQ runs validate the performance level of the relevance feedback base runs (Set A). (It's very easy to get large percentage increases for relevance feedback if the base case with no feedback is poor.) Second, the June MQ runs serve as a source of different relevant documents, since they presumably used different techniques that might not be possible to combine with relevance feedback. Third, they allow the stability of the entire evaluation process to be examined, since all the MQ runs were also fully evaluated (on 795 different topics) with the official separate running of a MQ-style evaluation.

Table 2 gives a comparison of the top groups of the Relevance Feedback and June MQ runs, evaluated using R-precision on the 31 Pool10 topics. Only the best run per group is given. MAP at 1000 documents would have been a preferable measure to use, but would

| Track | Run name | R-precision |
|---|---|---|
| RF | Brown.A1 | 0.2515 |
| RF | uogRF08.A1 | 0.2387 |
| MQ | txrun | 0.1935 |
| RF | UAmsR08PD.A1 | 0.1924 |
| MQ | indri25DM08 | 0.1896 |
| RF | HKPU.A1 | 0.1892 |
| RF | HitRF08.A1 | 0.1671 |
| RF | uams08bl.A1 | 0.1662 |
| RF | DUTIRRF08.A1 | 0.1633 |
| MQ | neumsfilt | 0.1617 |
| MQ | LucLpTfS | 0.1583 |
| RF | UIUC.A1 | 0.1582 |
| RF | THUFB.A1 | 0.1571 |

Table 2: Comparison of MQ and RF ad hoc (base case) runs on 31 topics

unfair to the MQ runs, since they often had less than 1000 documents retrieved after Set E documents were removed. R-precision was used since it induces system rankings very similar to MAP. As the table shows, the better base case relevance feedback runs were very competitive with the top MQ runs, with the best relevance feedback base run being 30% better than the best MQ run.

The June MQ results and comparison will eventually be explored fully in separate papers; the rest of the analysis in this paper will include the relevance feedback runs only.

## 4.3 MQ-style vs Pool10-style rankings

The first question to investigate is whether our evaluations are robust enough to be believed. There were 3 different poolings of documents

- Pool10 on 31 topics
- MQ-style for statMAP on 208 topics
- MQ-style for expectedMAP on 237 topics

The Pool10 style allows several different measures to be calculated, while the two MQ-style poolings are specific to one evaluation measure. Table 3 gives Kendall tau figures comparing 5 different rankings of the 118 relevance feedback runs (all sets of inputs).

The agreement between Pool10 MAP and the MQ-style statMAP and expectedMAP is reasonable, given the different topics it is based on. The rankings are closer than that between MAP and P(10), which are on the same topics, but much different than that of MAP and R-precision.

| Measure 1 | Measure 2 | Kendall tau |
|---|---|---|
| MAP | statMAP | 0.7977 |
| MAP | expectedMAP | 0.7610 |
| MAP | P(10) | 0.7483 |
| MAP | R-precision | 0.9210 |
| statMAP | expectedMAP | 0.8663 |
| statMAP | P(10) | 0.6340 |
| statMAP | R-precision | 0.7818 |
| expectedMAP | P(10) | 0.6039 |
| expectedMAP | R-precision | 0.7422 |
| P(10) | R-precision | 0.7479 |

Table 3: Kendall tau agreement between pairs of evaluation measures

If we consider runs to be tied in the rankings if the difference between them is less than 5% of the range of values of that measure, then a comparison between all pairs of runs in the rankings of MAP and statMAP shows that 5315 pairs agree in their order, 144 disagree, and 1009 involve tied runs. Thus the level of strong disagreements between MAP and statMAP is low.

# 5 Relevance Feedback results

Figures 1 through 4 give the results for most groups for their best set of runs. The MAP and P(10) plots are from the Pool10 evaluation, and the statMAP and expectedMAP plots are from their MQ-style evaluations.

It can be seen there is general agreement in the system rankings comparing the Pool10 and MQ-style evaluation. The P(10) evaluation has a bit more significant differences, as would be expected.

One striking feature about all the plots is that only a third or less of the systems get monotonic improvements as the amount of relevance information increased. Most systems increase most of the time, but not all the time. How much of this is due to systems not being well tuned, and how much is due to evaluation uncertainty (especially for those systems with less feedback effect), is unclear.

There is an enormous range in the base case values (Set A). Even with Set E relevance information, many groups do not achieve even the base case performance of the top groups. Are top groups like Glasgow, which has reasonably flat performance, already using the same sort of information they can get from explicit relevance feedback, and thus can't be improved? It remains to be seen what is happening here.

The P(10) plot gives a concrete interpretation of what the improvement due to relevance feedback is. Sabir, starting from a very low base case, went from an average of 1.5 relevant documents in the top 10, to 4.4 relevant documents in the top 10. Several other groups

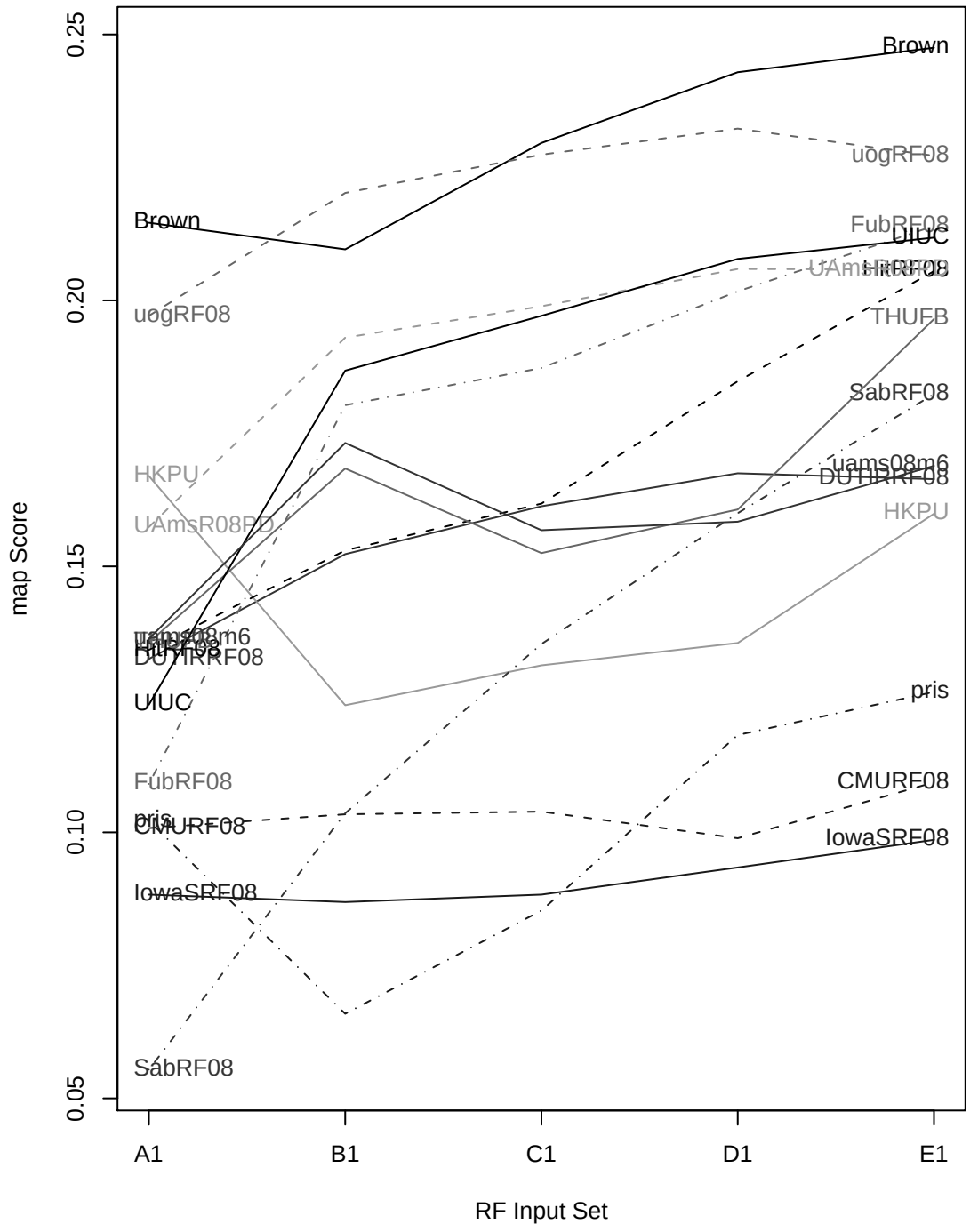


Figure 1: MAP scores for increasing relevance information (31 topics)

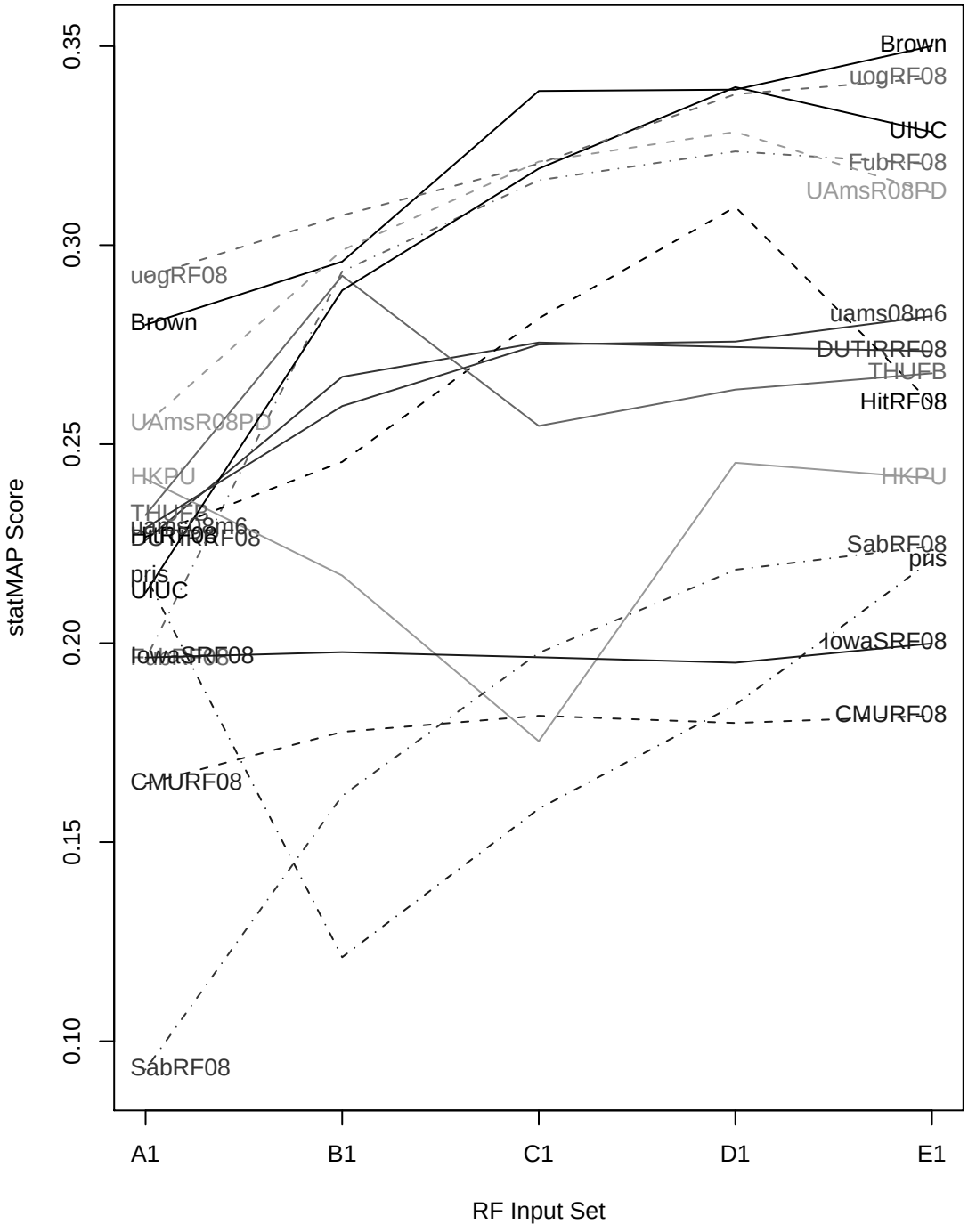


Figure 2: statMAP scores for increasing relevance information (208 topics)

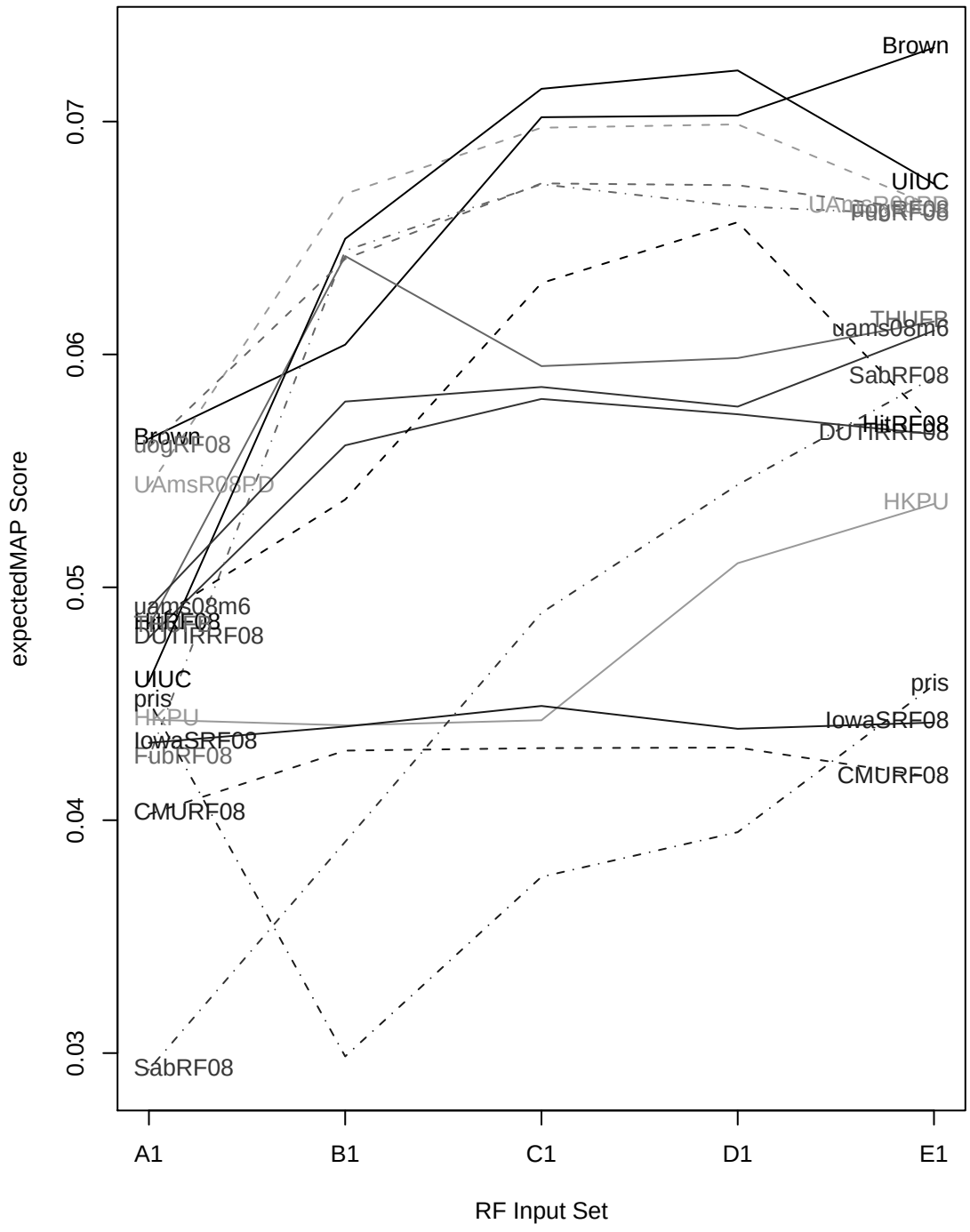


Figure 3: expectedMAP scores for increasing relevance information (237 topics)

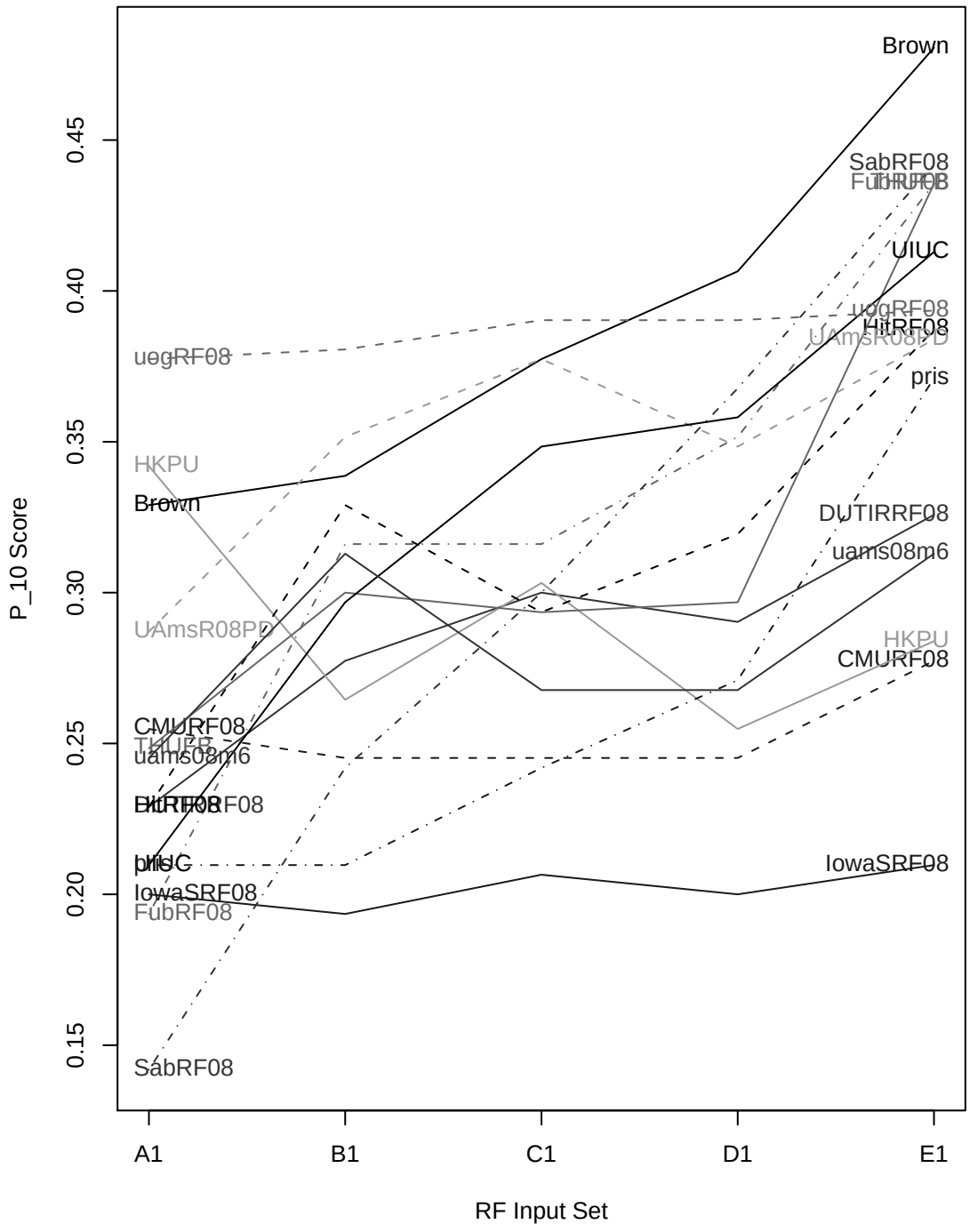


Figure 4: P(10) scores for increasing relevance information (31 topics)

| Residual Collection | Kendall tau with Set E Residual Collection |
|---|---|
| A | 0.9619 |
| B | 0.9048 |
| C | 0.9378 |
| D | 0.8667 |
| E | 1.0000 |

Table 4: Kendall tau agreement between rankings with different residual collections

gained 2 relevant documents in the top 10. The top P(10) group, Brown, gained 1.5 relevant documents. But the top base case group, Glasgow, stayed flat in the top 10 across the various sets (though their MAP scores show they got improvements later on in the retrieval.) It will be interesting to see whether groups like Sabir are getting their increases on exactly the same topics that Glasgow is already doing well on. The top base case groups are already retrieving 9 to 10 relevant documents on numerous topics; they can't improve their P(10) performance on those.

The results do indicate that determining an expected benefit from adding a specific amount of relevance information is going to be difficult to do. The systems varied greatly in whether they could take advantage of the additional information - detailed analysis at the system and topic level is going to be required.

# 6 Maximal residual collection results

One question discussed earlier was whether there can be more accurate comparisons of runs by minimizing the set of removed documents: if a comparison between 2 Set B runs is desired, then just remove the Set B documents that were used as input for those runs. This process severely limits the comparisons that can be made: only runs with the same removed set can be compared, and only the Pool10 evaluation can be used. However, it substantially increases the number of judged documents in the comparisons, since now the rest of the Set E documents that are not in Set B can be used (and judgments exist for all of them). Thus the accuracy of the comparison might be improved.

This can be tested by comparing the rankings of systems using the maximal residual collection (minimizing the removed documents) against the rankings of those systems after removing the Set E documents. This was done for the MAP measure for each of the A1, B1, C1, D1, and E1 sets of runs, with the Kendall tau results in Table 4.

The individual scores of systems were changed dramatically (for example, Brown went from MAP of .2146 to .3729), but the ranking of systems was not changed much. All of the movements in rank were minor. Thus the Set E residual collection seems to give enough information to accurately rank systems, and allows us to ignore the effects of general system tuning on past collections and judgments.

# 7 Research results of the participating groups

The participants in the track studied a wide variety of aspects of the relevance feedback algorithm. A number of them, especially among the top-performing groups, focused on basic language modeling approaches to combine pseudo-relevance feedback information with relevance information (Brown, CMU, Fondazione (FUB), Amsterdam (UAms), Iowa). In general, the pseudo-relevance information helped, both in the base case runs and the RF runs. Glasgow used pseudo-relevance information only for their strong base case performance.

Several groups looked at the use of non-relevant documents as well as relevant documents, particularly Fondazione (FUB) and Amsterdam-ILPS (uams), but also Amsterdam (UAms), Glasgow, and Sabir. Most found the non-relevant documents to be of little help, as has been found in the past. Amsterdam-ILPS got consistent small improvement, but except in the case of large amounts of non-relevant information (set E), estimating non-relevance information based on the entire collection worked better than using the judged non-relevant documents.

Illinois (UIUC) got nice performance and improvements by focusing on the balance between the original query and the positive relevance information on a per topic basis. They incorporated several techniques for each topic to estimate the importance of the feedback information based on topic and document analysis, with most of them yielding improvements.

Several groups looked at passages or term proximity in documents, including Hong Kong, Glasgow, and Iowa, with mixed success. Brown also looked at adjacency of terms in the topic itself. Glasgow in particular looked at the use of syntactic analysis to get document surrogates for use in expansion, but was not able to show improvements with that technique.

Tsinghua (THUFB) looked at document-document similarities directly instead of the query expansion and weighting approaches of the others. Direct document-document similarities was helpful, though clustering was not.

Term selection was a focus of two groups, Fondazione and RMIT, though most groups had to study it a bit. RMIT's result was the highly weighted terms may not be the best ones to use for expansion.

# 8 Conclusion

The TREC 2008 Relevance Feedback track attracted 15 groups who submitted 118 runs, which were evaluated with several different methodologies. The major emphasis for this year's track was examining how increased amounts of relevance information improved performance. Groups all ran with the same sets of relevance information which could then be compared both within the group and against other groups.

The results confirmed that relevance feedback consistently improves system performance. However, the amount of improvement remains very system dependent; a lot of work remains to understand why various systems reacted the way they did to the presence of additional information. Some preliminary investigations using overall system averages were presented here; it's clear that much more analysis at the topic level will be required to more fully understand relevance feedback.