TREC 6 High-Precision Track

Chris Buckley (SabIR Research Inc)

Abstract

The TREC High-Precision (HP) track compares systems on the simple "Real World" task of having users finding a few relevant documents as quickly as possible. Five groups are participating with each happening to emphasize different aspects of the retrieval process, from visualization to structured queries to relevance feedback. With so few groups participating in this inaugural run of the the track, no decisive conclusions can be reached. However, the indications are that simple approaches work very well.

Track Description

The High-Precision Track is a new track for TREC. It has a very simple short description: for each query, a user should find the best 10 documents possible within 5 minutes clock time. One realistic scenario corresponding to this task might be that your boss asks you for a quick report on some area and you need to find some information on the area fast.

There are no restrictions on the type of resources the user may use during this task other than

- Only one user per query per run (no human collaboration).
- The user and system can have no previous information about the query (eg, the system cannot have previously built a query dependent data structure.)

In particular, the users are allowed to make multiple retrieval runs, allowed to look at documents, allowed to use whatever visualization tools the system has, and allowed to use system or collection-dependent thesauruses, as long as they stay within the 5 minute clock time.

This track tests (at least) the effectiveness, efficiency, and user interface of the systems. The task provides a forum for testing many of the neat ideas in user interface and visualization that have been suggested over the years.

Unlike other interactive evaluations (for example, the TREC 6 Interactive task), no attempt is made to factor out user differences when comparing across systems. All users are assumed to be experts and equally proficient in use of their own system. This allows for fair comparison of systems, but implies that the absolute level of performance within the track will be better than the level obtainable from casual users. These are upper-bound interactive experiments.

The primary evaluation measure is precision at 10 documents. If less than 10 documents were submitted for a query, the missing documents were counted as non-relevant. *Relative precision* at 10 documents is also calculated. Relative precision gives the precision score relative to the maximum precision score possible. For example, if a query only has two relevant documents, then retrieving both in the top 10 will give a precision score of .20, as opposed to a relative precision score of 1.0. Relative precision is good for those evaluation environments where results are averaged over queries; it weights each query equally, as opposed to precision which considers queries with very few relevant documents as less important. A third experimental measure is unranked average precision, roughly corresponding to the average precision evaluations in other tasks (with most relevant documents having a precision of 0 since they are not retrieved).

Run	Precision	Relative Precision	Unranked Avg-Precision
Cor6HP3	.6020	.6298	.1021
otc3	.5800	.6020	.1067
uwmt6h0	.5720	.5977	.0902
uwmt6h1	.5680	.5834	.0982
Cor6HP2	.5660	.5820	.0786
uwmt6h2	.5640	.5951	.0997
otc1	.5500	.5700	.0973
Cor6HP1	.5440	.5564	.0799
otc2	.5440	.5587	.0916
pirc7Ha	.4260	.4384	.0574
pirc7Ht	.3980	.4163	.0766
DCU97HP	.3820	.4031	.0633
pirc7Hd	.3360	.3509	.0561

Table 1: High Precision Results - 50 Queries

In general, once averaging occurs

- Precision is biased against queries with fewer than 10 relevant documents.
- Relative Precision treats all queries equally.
- Unranked average precision strongly favors those queries with few relevant documents.

The same document set and 50 queries are used for the HP track as for the TREC-6 ad-hoc task. Each group could turn in up to 3 runs. NIST put all retrieved HP documents into the ad-hoc judgement pool; this added few documents to the pool since only 10 documents per query were submitted.

Track Results

Five groups submitted runs in the HP track, using a variety of approaches.

1. Cornell/SabIR: Used parts of the original topic as a query, judged documents, and used automatic relevance feedback.
2. Dublin: Visualization of individual documents (Document At A Glance)
3. OpenText: Submitted runs but did not attend conference or describe runs.
4. Queens: Submitted top 10 documents of automatic runs tuned for high precision (no users).
5. Waterloo: Faceted Boolean queries, choice of sets of documents, passages.

Table 1 gives the results for all runs in the HP track. The results are sorted in decreasing order of precision. Comparing the ordering of groups in each column, relative precision gives basically the same ordering as precision. Where there are swaps in ordering (e.g., between Cor6HP2 and uwmt6h2), we know that one system (in this case Cor6HP2) did comparatively poorly on those queries with less than 10 relevant documents. That is confirmed by looking at the third column which even more strongly favors queries with fewer relevant documents. (The last column is so strongly affected by queries with few relevant documents that it is not a reasonable evaluation measure. This measure will not be used in future HP evaluations, though it is valid in other environments where the system determines the number of retrieved documents.)

Given the preliminary nature of this track, there is not too much analysis possible. The Pirc runs were entirely automatic, and as would be expected, did not do well when compared against the other manual runs. Dublin's document-at-a-glance approach had serious efficiency problems and was done by a naive user. OpenText did well, but unfortunately has not ever described what they did! The Cornell/SabIR and Waterloo groups also did well; but each group thinks they can improve significantly now that they know where the critical points are.

Cornell/SabIR used the same basic approach for all 3 of their runs, as did Waterloo. The differences in the 3 results are due to the individual tweaks that each user added to the basic approach. For example, Cor6HP3 used much smaller initial queries than Cor6HP2 or Cor6HP1. Thus the system was able to finish the good complicated relevance feedback algorithms much more quickly than with the other users, and thus judgements were made on a better set of documents. It is clear that computational efficiency played a big part in the differences between runs.

Conclusion

The overall design of the High-Precision track at TREC 6 worked well. The participants all agreed that it was a good task that could be fairly evaluated and that they all learned substantial amounts about their system, able to improve their system for the TREC 6 task and getting ideas for future improvements. There were not enough full participants to be able to draw any conclusions about the relative value of the basic approaches. The preliminary results indicate that light user interaction (user only judges documents relevant or not) does as well as more complicated user interactions. However, any definitive decision on this remains for future TRECs!