The TREC 7 High Precision Track

Chris Buckley - SabIR Research, Inc.

Track Overview

TREC 7 is the second year the High-Precision (HP) track has been run. It is an attempt to perform a task that is much more closely related to real-world user interactions than the ad-hoc or routing task. The goal is simple: a user is asked to find 15 relevant documents in 5 minutes. No other restrictions are put on the user (other than no prior knowledge of the query, and no asking other users for help). Official evaluation is simply how many actual relevant documents were found among the 15 documents supplied by the user, modified slightly for those queries with fewer than 15 relevant documents in the collection (Relative Precision at 15 documents).

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 only changes in the rules from the TREC 6 track are to raise the number of relevant documents required to 15 instead of 10, and to forbid cutting and pasting of the original query. This latter change requires the participants to type in the query, and makes the task fairer for those groups for whom cutting and pasting would not give a query in the proper form. It also has the side effect of making the task more difficult since reading and typing parts of the query might take 30 seconds (10% of the available time).

High-Precision Results

Four groups participated in the High-Precision track with a total of 7 runs.

- Cornell University/Sabir Research (3 runs)
- University of Waterloo (2 runs)
- Australian National University (1 run)
- CUNY-Queens: PIRCS system (1 run)

All four groups used basically the same retrieval approach as they used for their ad-hoc runs. In all cases, the majority of the user time was spent just judging documents that were sequentially given to the user by the system, with occasional query reformulation. All groups allowed the user to reformulate the query by hand; Cornell also used automatic relevance feedback to expand the query based on terms from seen relevant documents.

All four groups did quite well on the task. The CUNY group had a slight misunderstanding of the task that adversely affected their score.

Run	Precision	Relative Precision	Num relevant
Cor7HP3	.5853	.5967	439
Cor7HP2	.5813	.5920	436
Cor7HP1	.5787	.5909	434
uwmt7h1	.5693	.5772	427
uwmt7h2	.5373	.5467	403
acsys7hp	.5120	.5295	384
pircs8Ha	.4773	.4839	358

Agreements with TREC Assessors.

One important question is how the users agree with the official TREC relevance judgements. If the HP track is to have meaning, the disagreement between user interpretation of relevance to a query, and the official assessor interpretation can not dominate the results. Both Cornell and Waterloo studied how their judgements agreed with the official NIST judgements. The Waterloo high-precision runs were a subset of their manual ad-hoc runs and they didn't separate out their high-precision figures, but overall they agreed with the Cornell figures.

For the three Cornell runs, Table 1 gives the total number judged relevant, possibly relevant, and non-relevant for each user, for both the TREC-assessor judged relevant documents and the TREC-assessor judged non-relevant documents. For example, Cornell User 3 judged 290 documents (159+131) relevant or iffy that the official assessors had judged non-relevant.

Run	TREC judged Rel			TREC judged NonRel			Overlap
	UserRel	Iffy	NonRel	UserRel	Iffy	NonRel	(Iffy=rel)
Cor7HP1	315	170	51	79	181	448	61%
Cor7HP2	396	73	36	115	128	444	63%
Cor7HP3	374	100	84	159	131	674	56%

Table 1: Cornell High-Precision User-assessor consistency (50 queries)

The last column gives the overlap on judgements of relevant documents. Overlap is defined as the intersection of the relevant judgements divided by the union of the relevant judgements, and is the desired measure for how well judgements agree. (Early TREC studies used other measures that are dependent on the number of total number of documents judged. These measures might have been reasonable for those particular studies since the total number of documents was fixed, but the measures are *not* valid for general use, despite the fact that others have adopted them.)

The great majority of the disagreements are the users considering documents relevant that the assessor considered non-relevant. In fact, consider the 15 queries with lowest overlap for each of the three users; for all 45 queries the user has looser criteria than the assessor. This is to be expected, since the assessor as the originator of the query can easily have in mind a stricter query than made it to the topic description. For example, in query 375 "hydrogen energy", the assessor obviously did not want hydrogen fuel for car engines, though that wasn't clear from the topic. The three users marked

a total 50 documents as relevant or iffy that were not relevant. Query 363 "tunnel disasters" was another with major disagreements (36 documents).

The disagreements in the other direction are rarer and a bit less obvious. For example, query 377, "cigar smoking", had the most disagreements, with 15 total assessor relevant documents being marked non-relevant by the three users.

The overall level of disagreement between assessor and users is unfortunately high. The overall level of performance is being strongly affected by agreement with assessor, rather than intrinsic performance.

Difficulty of Task.

One of the ways of telling how easy or difficult the TREC 7 HP task is, is to look at the queries for which the users did not find 15 documents that they thought were relevant. Table 2 gives the number of documents that are included in the final submitted retrieval without being judged for the three Cornell runs. There will be unjudged documents only if the user did not find 15 relevant or iffy documents after 5 minutes.

Run	num docs unjudged	num queries with unjudged	num unjudged rel docs
Cor7HP1	122	24	12
Cor7HP2	139	25	20
Cor7HP3	84	17	13

Table 2: Cornell Unjudged Retrieved Documents

Half or less of the 50 queries have any unjudged documents at all for all three Cornell users. This includes queries for which there were fewer than 15 relevant documents in the collection. This implies for the majority of the queries, the only evaluation differences are due to disagreement with assessors rather than effectiveness of system. Combined with the high disagreement between users and assessors, the conclusion must be reached that the task is too easy.

Put another way, on average for Cornell User 2, of 15 documents returned per query

- 8.8 were agreed relevant
- 2.4 were agreed non-relevant
- 3.5 had relevance disagreements

The disagreements are more important than non-found documents. Again, this suggests the task was too easy and didn't stress the users and system enough.

Timing Evaluation.

Cornell, Waterloo, and ANU kept track of not only what each user document judgement was, but when it occurred (though ANU only has figures for 39 out of the 50 queries). Thus we can analyze the time performance of each user, and hopefully develop time-based evaluation measures that reflect the power and efficiency of systems.

The most obvious fact to look at is when the relevant documents were retrieved. Figure 1 gives the number of relevant documents retrieved during each 5 second timeslice for Cornell User 1, on average for 50 queries. The number of retrieved relevant starts off at 0 for the first 20 to 50 seconds as the user reads and types in the query. Then it steadily increases for the next minute or so and then starts slowly decreasing up until the 5 minute point is reached. There's a big hump at 300 seconds as the 15

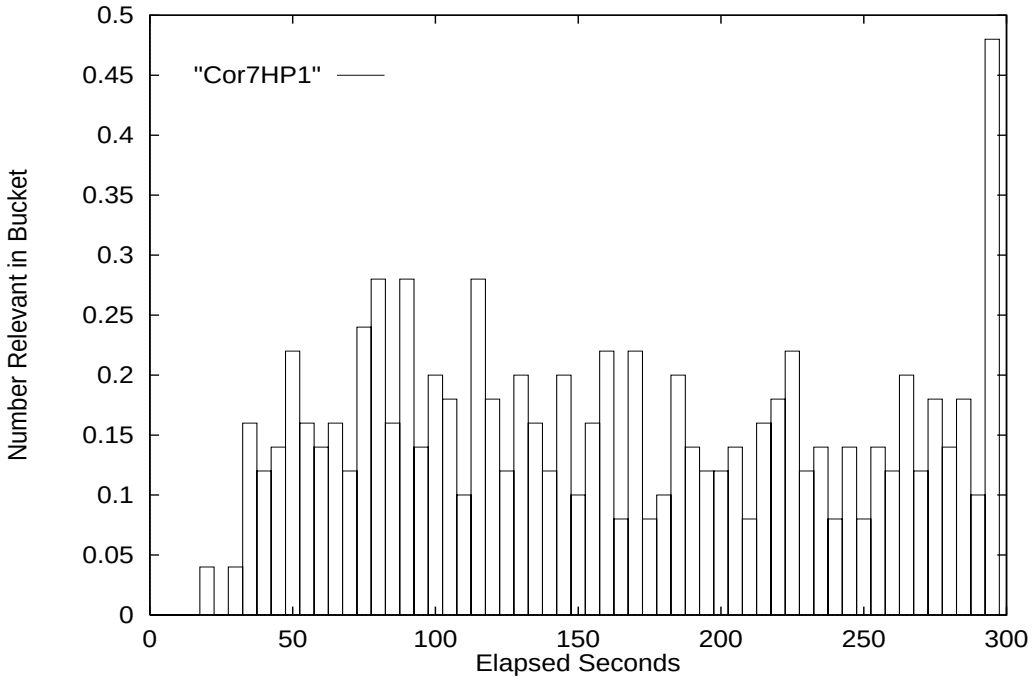

Figure 1: Average Relevant Retrieved per 5 second Timeslice over 50 Queries

documents to be returned get filled in with unjudged documents. In the normal course of retrieval , these documents would be judged over the next few buckets.

This graph is actually evidence against the conclusion reached earlier that the task was too easy. The rate at which relevant documents are being added close to 300 seconds is still substantial. The previous evidence indicates it can't go on for much longer, and that less than half of the queries are still active. However, there is no sudden drop-off as there would be if this particular run finds too many relevant documents.

Figure 2 compares the three Cornell users on a typical single query, Query 366. The measure being plotted is precision at 15 documents. More generally, the time-precision measure is defined as
Time-precision = num-relevant-retrieved-so-far / total-session-retrieved

User 2 typed in a shorter query so started judging documents earlier than the others. User 2 maintains a lead up until 180 seconds, when User 1 takes over. Then at 240 seconds, User 3 takes the lead for the last minute.

For this particular query, it is clear that User 3 has the best end result (precision after 5 minutes). But it is also clear that User 2 and possibly User 1 have better sessions: they find relevant documents sooner during the first 4 minutes.

Figure 3 gives the same comparison except on the average of all 50 queries. Once again, User 2 has the lead for most of the session up until the very end when User 3 takes over. For most of the session, User 2 is about 10 seconds ahead of User 3 and 20 seconds ahead of User 1. Again, User 3 has the best end result, but User 2 had the best session.

Other evaluation measures give the same overall results. For example, Unranked Average Precision at 15 documents is given in Figure 4. The curve is almost identical.

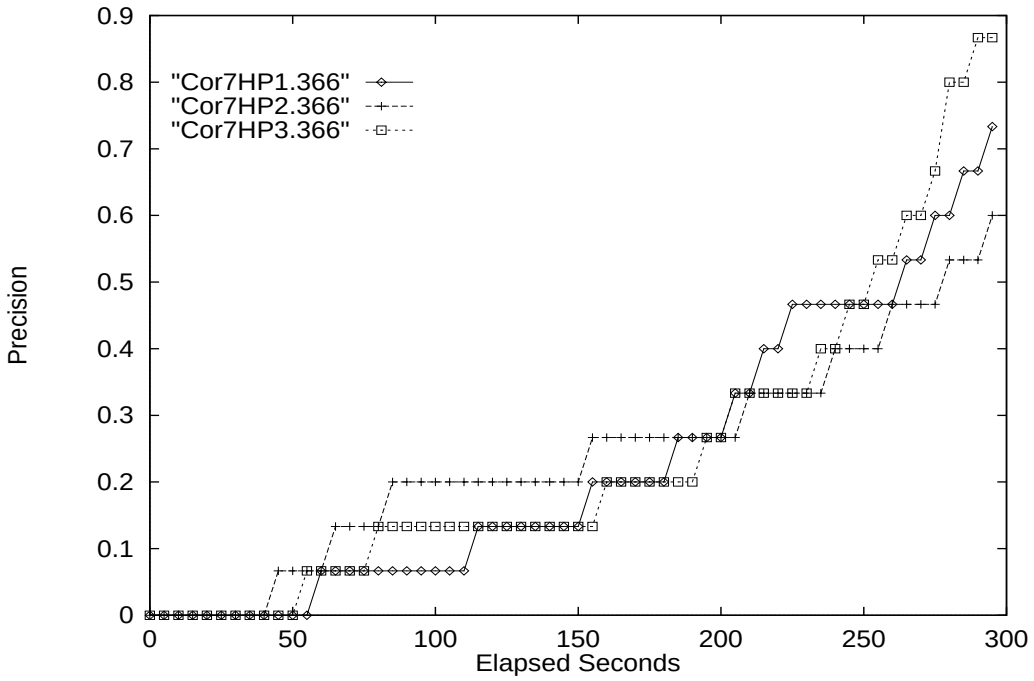

Figure 2: Time-Precision (at 15 Documents) vs. Time for Query 366

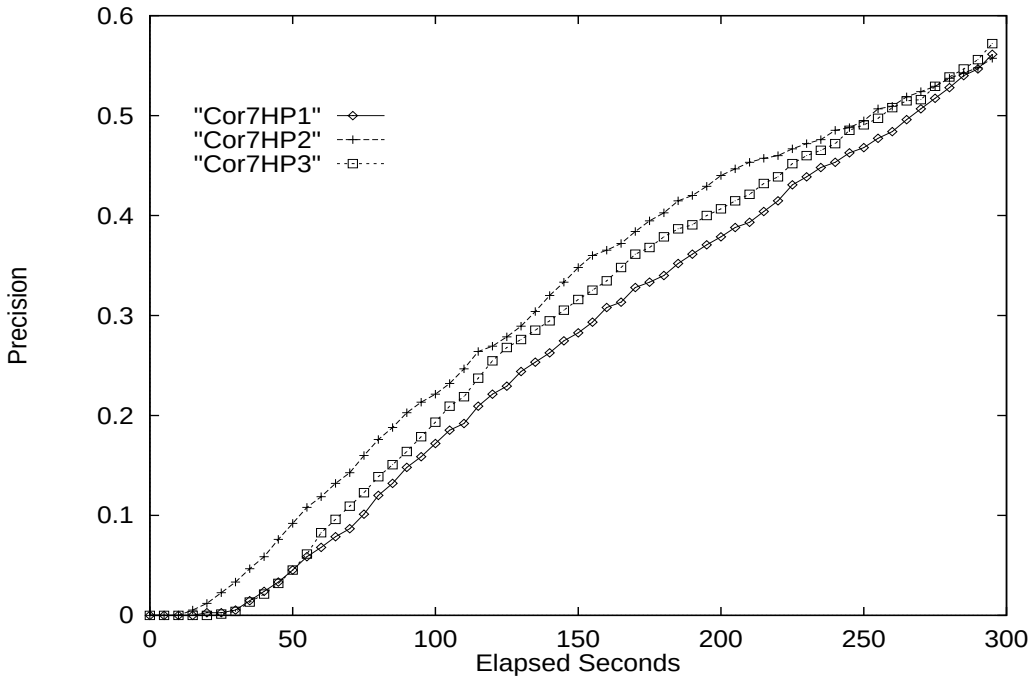

Figure 3: Time-Precision (at 15 Documents) vs. Time over 50 Queries

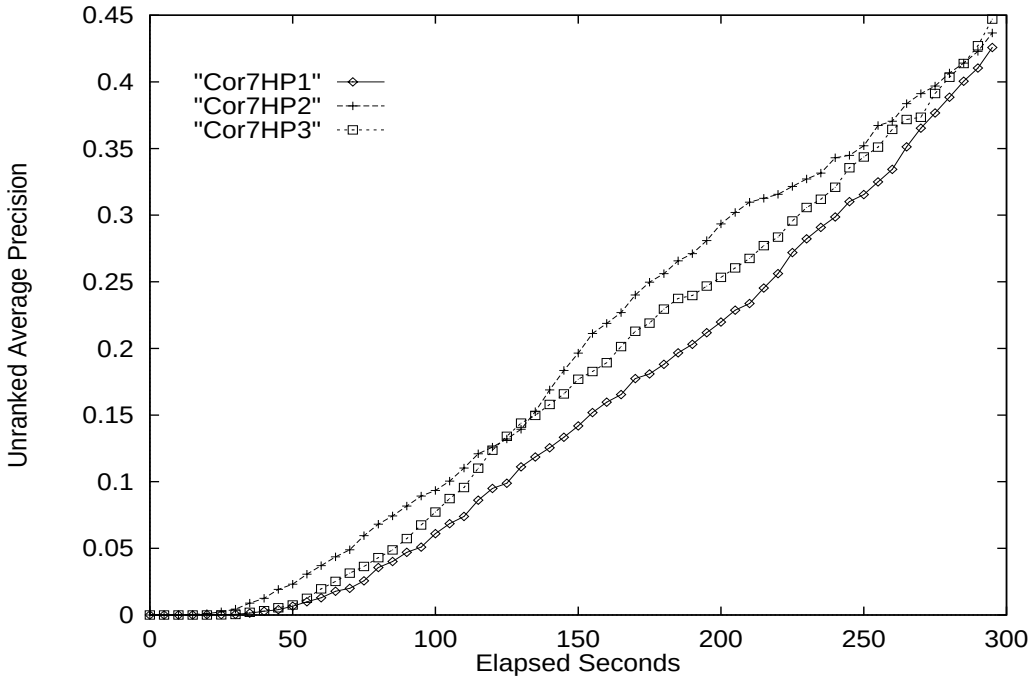

Figure 4: Unranked Average Precision vs. Time over 50 queries

One different evaluation measure is Utility(1,-1,0,0) in Figure 5. This measures increases by 1 when a relevant document is retrieved and decreases by 1 when a non-relevant document is retrieved. It is a poor evaluation measure for the HP task. It is dominated by the retrieved non-relevant documents; i.e., those documents for which user and assessor disagree on relevance. None-the-less, the results are informative.

User 2's lead is even more substantial (remember User 2 has the most accurate judgements as measured by agreement with assessors). But what is very interesting is how the plots for User 2 and User 3 flatten out over the last 2 minutes. For every relevant document being added, a non-relevant document is being added. This may indicate more disagreements occur late, or maybe there is a natural stopping spot late. Further study is needed, especially since the handling of "iffy" documents may be partly responsible for the effect.

All of these time-based measures and graphs suggest that a reasonable evaluation measure for an entire session is the area under each plot, much in the same way as the area under the recall-precision curve is a good single measure (this is "average precision"). Table 3 gives three such measures, corresponding to the three different plots seen above. As expected, for all 3 session measures, User 2 has a substantial $(6\% - 8\%)$ lead over User 3 and even more over User 1. It can certainly be argued that this evaluation approach is a better approach than the official measure used for the track (precision after 5 minutes.)

Run	Average Precis	Average UAP	Average Utility(1,-1)
Cor7HP1	.2726	.1590	3.997
Cor7HP2	.3104	.1934	4.606
Cor7HP3	.2901	.1780	4.287

Table 3: Timing Evaluation

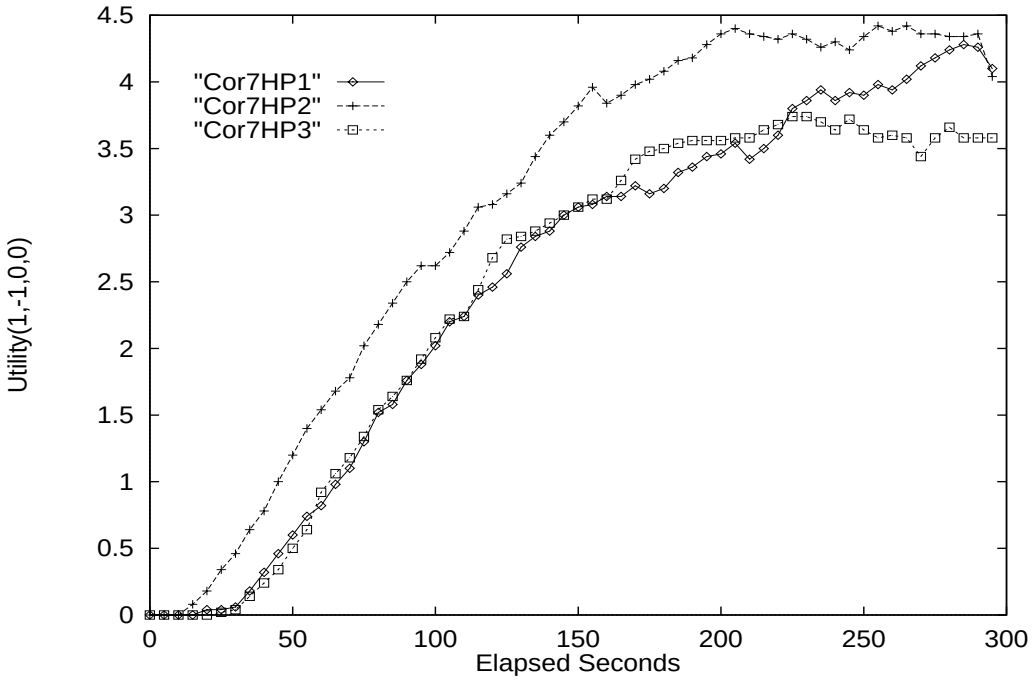

Figure 5: Utility(1,-1,0,0) vs. Time over 50 queries

These session evaluation measures can be extended to work on any time-based retrieval. It would be very interesting to apply these measures to the standard Manual portion of the ad-hoc task. Perhaps in the future, we can request that timing figures be optionally supplied, perhaps as the iteration field, to Manual submissions. There are still open questions regarding these measures. A couple that immediately spring to mind is how sensitive they are to starting time, and to size of time-slice. However, they still seem to offer a hope at bringing efficiency into evaluation of manual systems and sessions.

Note that the latest copy of trec_eval is in pub/smart/trec_eval.7.0beta.tar.gz on ftp.cs.cornell.edu and includes all the measures discussed here plus many others, though perhaps not in their final form (for instance, the timing information is assumed to be in the "sim" field but will probably be moved.)

Conclusion

The High-Precision track of TREC 8 was an attempt to evaluate a realistic user-oriented task, namely finding relevant documents quickly. All the groups did well, coming very close to the limits of performance that relevance judgement disagreements allow. This suggests that the task was probably too easy and didn't stress the user and systems enough.

It was unfortunate that no group used any really innovative user interface, such as looking at clustered retrieved documents. Instead, all groups took the approach of trying to judge as many good individual documents as quickly as possible. The experimental setup allows for much more interesting comparisons than were done this year.

After the experiment, we looked in-depth at methods of analyzing and evaluating time-dependent retrieval sessions. We came up with several new evaluation measures that seem to capture the essentials of what a session evaluation of manual retrieval should capture. These approaches may be quite useful outside of the High-Precision track, perhaps to evaluate timed Manual retrieval.