SIGIR Workshop on Interactive Retrieval at TREC and Beyond

William Hersh, Oregon Health Sciences University, hersh@ohsu.edu
Paul Over, National Institute of Standards and Technology, over@nist.gov

Attendees: Nick Belkin, Aurelien Benel, Pia Borlund, Matthew Carey, Efthimis Efthimiadis, Ayse Goker, Donna Harman, Dajing He, Bill Hersh, Peter Ingwersen, Ray Larson, Anton Leuski, Robert McArthur, Gheorge Muresan, Paul Over, Mark Sanderson, Ross Wilkinson, Byoung-Tak Zhang, Joe Zhou

The annual Text Retrieval Conference (TREC) has become a major event in information retrieval (IR) research. Over the last few years, its main focus has changed from the common ad hoc and routing tasks to tracks focused on special interests, such as cross-language retrieval, question answering, and large database/Web retrieval. In each of the tracks, a common task and test collection is developed each year for all participating groups to undertake. One problem the tracks have is a paucity of face-to-face meetings, since the only certain time each track actually meets is at the annual TREC meeting. For many tracks, this time is too early to work out all the details of the coming year's track, and most additional work is done via email.

The TREC Interactive Track has been operating since TREC-3, and has developed a small but steady group of participants over the last few years. A summary of work by the major participants of the track will be published in an upcoming special issue of *Information Processing & Management*. One of the problems for the track has been a lack of time for long-range planning. We therefore decided during the TREC-8 conference (where we were planning for TREC-9) that we would put on a workshop at SIGIR 2000 with the goals of planning the future of the track, encouraging new participants to take part in the track, and providing perspective on our past work.

The workshop was attended by 19 individuals - a mixture of those previously or currently active in the track, newcomers, and others, who have not participated in TREC but are active in the area of interactive retrieval evaluation. The workshop began with brief presentations by four current participants on what has been learned and how we might improve the track. These were followed by presentations from five researchers who have not participated in the track. They provided their views of the track, its weaknesses, and where it should be headed. There followed a brainstorming session focussed on completing the following sentence, "We can learn more in TREC about how to build better interactive IR systems if we …" The remainder of the workshop was devoted to choosing from the ideas on the table and exploring the pros and cons of each.

Bill Hersh led off the presentations by current participants by describing what we have learned and not learned in the TREC interactive track. His results, while not definitive, suggest that what developers and users want may not provide the anticipated benefit. For example, his work suggests that system features shown in batch studies to be beneficial do not necessarily hold up with real-world searcher tasks (e.g., his paper published in the conference proceedings). Others (e.g., Belkin, Larson) have shown that system features requested by users do not always benefit them. He was followed by Nick Belkin, Ross Wilkinson, and Ray Larson, each of whom provided his perspectives on past results and where he would like to see the track go.

These presenters also described some of the basics things (re)learned about the logistics of human studies in the track. Interactive experiments are, in general, relatively expensive in terms of time and money compared to batch experiments. Users do not always follow instructions and their departures from carefully planned procedures are not easy to compensate for within the short TREC cycle. This short cycle also makes experiments very susceptible to hardware problems and software errors, i.e., you can't just fix a bug and quickly do another run. It also rushes us to move on to the next study before fully analyzing the results of the last one.

The presenters pointed out that there is a great deal that we have *not* learned. For example, our small number of tasks, queries, and users limit the generalizability of our findings. We also are not sure whether our methods are testing the right system features with the right tasks in realistic enough situations.

After the four track participants' presentations, five other researchers (Pia Borlund, Ayse Goker, Peter Ingwesen, Mark Sanderson, and Efthimis Efhimiadis) gave brief talks describing their research goals and how the interactive track could benefit their work. Some speakers were critical of the overall TREC experimental model, arguing that the pure laboratory approach with questions generated and judged by evaluators at NIST could not capture the reality of interactive searching. The counter-argument was that the consistency afforded by controlled experiments allowed experimental observations that could not be obtained with users generating their own queries and judgments.

There was also some discussion on what we might like to learn from controlled interactive searching experiments. The following list was developed:

- How to evaluate system features desired by developers and users
- How to cut through the hype of developers, academic or commercial
- How to develop an experimental methodology enabling us to answer research questions about IR systems in the hands of real users

This was followed by discussion on the best way to get there. It was thought that we need to:

- Develop appropriate research questions
- Figure out which can be answered by TREC Interactive Track
- Modify track to answer questions without losing its commonality

There was also sentiment that we need to take advantage of the strengths of TREC and there was a call for a return to comparability of systems across sites. The workshop agreed generally on the need for the full range of studies to understand user interaction in IR, but realized that TREC offers a unique environment with constraints but also strengths. It is clear that the use of common collections, tasks, and queries lends itself to experimental rather than observational studies. So the challenge is how can we maximize range of studies within this framework.

After the presentations and the brainstorming, the discussion moved toward making some specific recommendations for improving the track starting with TREC-10. The first suggestion that achieved some consensus was to consider putting the track on a two-year cycle. New tasks, collections, and queries would be introduced only every other year. Participants would report on

preliminary data and experiences one year and then full results the following year. It was hoped this change would provide more time for analysis of results.

The next suggestion was motivated in part by dissatisfaction with the age of the document collections used by the track and a desire to support search tasks that subjects would be more naturally motivated to carry out. This led to realization that there was a clear desire to study Web searching, though there were many questions associated with this choice that would require investigation, e.g., would we want to search the "live Web" or have the control over some large but finite extracted portion? Related to Web searching, Nick Belkin described some results from a graduate student of his who found four major uses of the Web by people in university settings that we could adapt for the track:

- Questions about personal health
- Buying a given item
- Planning travel to a specific place
- Finding information relevant for a project

The workshop decided to suggest that for the next year, the track should develop Web search tasks - observational studies, initially, but with experimental studies in the second year, perhaps using a Web corpus extracted during the observational studies. The workshop agreed the track should allow researchers to look at as many different types of searchers as possible. Among the groups we know we have available are:

- University students
- Medical and nursing students
- NIST assessors

There was also some discussion on how to collect data more completely and consistently. Robert McArthur of DTSC in Australia described a Microsoft Internet Explorer plug-in developed at his center that captures user interface events. The author may be willing to share it, which will allow more detailed logging than is currently done.

In summary, by the end of the day consensus had been reached about the following suggested points to be fleshed out and incorporated into a set of proposed guidelines for discussion at TREC-9 and on the track email discussion list:

- Relieve some of the pressure on participants by running the track on a two-year cycle with interim results reported after the first year; a new task only every other year
- Move the search task closer to everyday searching where, for example, duplication of information, recency, authority, etc. matter, by using live Web data and deal with the implications of its heterogeneous and dynamic nature for evaluation
- Define Web search tasks in four domains chosen based on popular Web usage - tasks experimental searchers should be able to identify with based on a simple cover story:
    - finding personal medical information on a given subject
    - buying a given item
    - planning travel to a given place
    - collecting material for a project on a given subject

- For at least for the first two-year cycle (TREC-10/11) allow participants to undertake mainly observational studies during the first year, but designed to support metrics-based comparison of systems by the end of the second year
- Aim for about 25 statements of information (topics) need rather than the six to eight the track has been using, so that the scope of conclusions with respect to topics can be improved. Adjust the experimental design to accommodate the increase while any given searcher uses only a practical subset.

There was also agreement on the desirability of finding resources to enable participants to have more protected research time, research assistants, and hardware resources to do experiments. One idea that we batted around at the TREC-8 track workshop and this workshop was trying to write a grant that would enable us to get such funding. Since most of the institutions involved in TREC are Internet2 institutions, one possibility would be to propose the use of Internet2 for cross-site searching. Among the possible funding initiatives we could consider are:

- The NSF International Digital Library Initiative
- The NSF Information Technology Research (ITR) Initiative

The discussion begun at this workshop will be carried out over email on the Interactive Track listserv, trec-int@ohsu.edu. Those interested in joining the list should contact Bill Hersh by email at hersh@ohsu.edu. The TREC-9 Interactive Track web page can be found at www-nlpir.nist.gov/~projects/t9i. Information about TREC in general is available at the TREC website (trec.nist.gov).