TREC-8 Interactive Track

William Hersh
hersh@ohsu.edu
Division of Medical Informatics & Outcomes Research
Oregon Health Sciences University
Portland, OR 97201, USA

Paul Over
over@nist.gov
Natural Language Processing and Information Retrieval Group
Information Access and User Interfaces Division
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA

April 5, 2000

For TREC-8 the high-level goal of the Interactive Track remained the investigation of searching as an interactive task by examining the process as well as the outcome. To this end a common experimental framework was designed with the following features:

- an interactive search task - a form of question answering
- 6 topics - brief statements of information need
- 12 searchers - a minimum, more if possible
- a document collection to be searched: Financial Times of London 1991-4 with 210,158 articles totaling 564 megabytes.
- a required set of searcher questionnaires
- 5 classes of data to be collected at each site and submitted to NIST
- 3 summary measures to be calculated by NIST for use by participating groups

The experimental design allowed groups to estimate the effect of their experimental manipulation free and clear of the main (additive) effects of searcher and topic and it was designed to reduce the effect of interactions, e.g., searcher with topic, topic with system, etc.

In TREC-8 the emphasis was on each group's exploration of different approaches to supporting the common searcher task and understanding the reasons for the results they get. No formal coordination of hypotheses or comparison of systems across sites was planned, but groups were encouraged to seek out and exploit synergies. Some groups designed/tailored their systems to optimize performance on the task; others simply used the task to exercise their system(s). Groups from the following institutions took part: New Mexico State University at Las Cruces, Oregon Health Sciences University, Royal Melbourne Institute of Technology/CSIRO, Rutgers University, Sheffield University, the University of California at Berkeley, and the University of North Carolina at Chapel Hill. A total of 936 searches were performed as part of the experiments.

In addition to running its experimental system(s), each participating site chose a control system appropriate to the local research goals. The control and the experimental systems were to be provided with equal computing resources within a site but not necessarily the same as those provided at other sites.

Six of the 50 topics created by NIST for the TREC-8 adhoc task were selected and modified for use in the interactive track by adding a section called "Instances" and removing the "Narrative." The six topics were entitled "tropical storms" (408i); "Cuba, sugar, imports" (414i); "declining birth rates" (428i); "robotic technology" (431i); "tourism, increase" (438i); and "tourists, violence" (446i).

Each of the six topics described a need for information of a particular type. Contained within the documents of the collection to be searched were multiple distinct examples or instances of the needed information. Here is an example interactive topic.

> Number: 408i
>
> Title: tropical storms
>
> Description: What tropical storms (hurricanes and typhoons) have caused property damage and/or loss of life?
>
> Instances: In the time alloted, please find as many DIFFERENT storms of the sort described above as you can. Please save at least one document for EACH such DIFFERENT storm. If one document discusses several such storms, then you need not save other documents that repeat those, since your goal is to identify as many DIFFERENT storms of the sort described above as possible.

The task of the interactive searcher was to save documents, which, taken together, contained as many different instances as possible of the type of information the topic expressed a need for - within a 20 minute time limit.

Searchers were encouraged to avoid saving documents which contribute no instances beyond those in documents already saved, but there was no scoring penalty for saving such documents and searchers were to be told that.

Each searcher performed six searches on the document collection using the six interactive track topics in a pseudo-random order. Each searcher performed 3 searches on one of the site's systems and then 3 on the other to avoid the extra cognitive load of switching systems with each search. Instructions on the task preceded all searching and a system tutorial preceded the first use of each system. In addition, each searcher was asked to complete a questionnaire, prior to all searching, after each search, after the last search on a given system, and after all searching was complete. The detailed experimental design determined the pseudo-random order in which each searcher used the systems (experimental and control) and topics.

Six sorts of data were collected for evaluation/analysis (for all searches unless otherwise specified) and are available from the TREC-8 Interactive Track web page (www-nlpir.nist.gov/projects/t8i/t8i.html). The types of data are sparse-format data - list of documents saved and the elapsed clock time for each search; rich-format data - searcher input and significant events in the course of the interaction and their timing; searcher questionnaires on background, user satifaction, etc.; a full narrative description of one interactive session for topic 408i; and any further guidance or refinement of the task specification given to the searchers.

Only the sparse-format data were evaluated at NIST to produce a triple for each search: instance precision and recall (these as defined in the next section) and elapsed clock time.

Evaluation by NIST of the sparse-format data proceeded as follows. For each topic, a pool was formed containing the unique documents saved by at least one searcher for that topic regardless of site. For each topic, the NIST assessor, normally the topic author, was asked to read all of the documents from the pool, determine the set of possible instances, record which instance(s) occurred in which document(s). Then for each search (by a given searcher for a given topic at a given site), NIST used the submitted list of selected documents and the assessor's instance-document mapping for the topic to calculate:

- the fraction of total instances (as determined by

the assessor) for the topic that are covered by the submitted documents (i.e., instance recall)

- the fraction of the submitted documents which contain one or more instances (i.e., instance precision)

The third measure, elapsed clock time, was taken directly from the submitted results for each search.

Comparison of systems across sites is not supported by the experimental design, so comparisons presented here are between systems within a given site. Still, a general theme running through the results was that there was little difference between each group's experimental and control systems even when they differed by the presence of techniques shown to be effective in non-interactive experiments in the past. This suggests these techniques, such as relevance feedback, Okapi weighting, document summarization, and greater control over search terms and Boolean operators, do not show benefit in the instance recall task. Only further research can show whether this is true or due to a lack of experimental power. The actual results obtained by each group are summarized in the following paragraphs.

- New Mexico State University looked at whether users could find relevant information with a user interface for viewing retrieval results that showed query term occurrence and distribution along with extracted names of people and locations shown in document surrogate lists and summaries. Their results showed no difference in instance recall between the two systems. However, 11 of 12 users reported that they liked the summary display better than the full text control. However, this preference did not match performance. Of note was that users viewed more documents in the full text condition.

- Oregon Health Sciences University used the interactive track to assess whether batch and user evaluations give the same results. Batch experiments with TREC-6/7 data showed substantial differences for various weighting schemes, and particular benefit for Okapi. User experiments showed no such comparable benefit. For more information see the site report (Hersh et al., 2000).

- Royal Melbourne Institute of Technology-CSIRO tested the hypothesis that by allowing the user control over the organization of the information, and the selection of documents using the organization, the user would find a better set of documents to view, and hence achieve a better coverage of aspects. Their control system featured three windows with a list of document titles, one document displayed, and a saved instances window. The experimental system replaced the list of document titles with a window containing a list of categories and documents clustered therein. The categories were derived from WordNet. Results showed that using the categorized interface, users read more documents, saved the same number of documents, and saved more aspects, but with less accuracy. User satisfaction did favor the categorized system. For more information see the site report (Fuller et al., 2000).

- Rutgers University compared two different techniques for supporting query reformulation by term suggestion in interactive IR: user-controlled relevance feedback (RF) and system-controlled Local Context Analysis (LCA). Their results showed that LCA did not perform better, but was easier for the user. Effectiveness and usability were the same for each system. In LCA mode, more terms were suggested than in RF and more suggested terms were used in the queries, which were equally long in both systems. Thus users had to do less (cognitive) work in LCA. The authors speculated that if LCA terms were "better," then maybe the approach would be more effective, usable, and preferred. For more information see the site report (Belkin et al., 2000).

- Sheffield University focused on searching behavior and user perception of an experimental retrieval task assessing the impact of document ranking, best-passage retrieval, and a query expansion facility. The experimental setting used two versions of Okapi, one with relevance feedback and one without. Their findings showed that while user outcomes were the same, search confidence was positively associated with the

number of instances retrieved. For more information see the site report (Beaulieu, Fowkes, Alemayehu, & Sanderson, 2000).

- University of California, Berkeley assessed new features added to its Cheshire II experimental system, in particular the Boolean NOT capability and new ways for navigating results and selecting relevant items. Users achieved the same instance recall as they did with the previous system. For more information see the site report (Larson, 2000).

- University of North Carolina found no difference among various levels of relevance feedback. For more information see the site report (Yang & Maglaughlin, 2000).

These results place an imperative on continued user-oriented evaluation. While non-interactive evaluation will continue to have its role, such as in assessing the feasibility of new algorithms and approaches and the parameterization of system features, interactive experiments must verify that new system advances can be used with their intended beneficiaries, real users. Just because users and user studies are unpredictable as well as resource-consuming, this does not mean we should avoid them.

Since the interactive track has focused on the instance recall task for three years running, a growing consensus of participating groups prefer to assess different retrieval tasks and documents. Next year's track will likely move to more of a question-answering approach using data from the Web track. The participants also hope to explore specific aspects of the interactive retrieval task. For example, future experiments might decompose the overall task into pieces, such as query composition or document selection. Likewise, there is a desire to base experiments on sound underlying models, such as those of the user, the task, and the typology of information needs. Future discussion will ensue on the track listserv (trec-int@ohsu.edu).

References

Beaulieu, M., Fowkes, H., Alemayehu, N., & Sanderson, M. (2000). Interactive Okapi at Sheffield - TREC-8. In E. M. Voorhees & D. K. Harman (Eds.), *The Eighth Text REtrieval Conference (TREC-8).* Gaithersburg, MD, USA.

Belkin, N. J., Head, J., Jegn, J., Kelly, D., Lin, S., Park, S. Y., Cool, C., Savage-Knepshield, P., & Sikora, C. (2000). Relevance Feedback versus Local Context Analysis as Term Suggestion Devices: Rutgers' TREC-8 Interactive Track Experience. In E. M. Voorhees & D. K. Harman (Eds.), *The Eighth Text REtrieval Conference (TREC-8).* Gaithersburg, MD, USA.

Fuller, M., Kaszkiel, M., Kimberley, S., Zobel, C., Justinand Ng, Wilkinson, R., & Wu, M. (2000). The RMIT/CSIRO Ad Hoc, Q&A, Web, Interactive, and Speech Experiments at TREC 8. In E. M. Voorhees & D. K. Harman (Eds.), *The Eighth Text REtrieval Conference (TREC-8).* Gaithersburg, MD, USA.

Hersh, W., Turpin, A., Price, S., Kraemer, D., Chan, B., Sacherek, L., & Olson, D. (2000). Do Batch and User Evaluations Give the Same Results?: An Analysis from the TREC-8 Interactive Track. In E. M. Voorhees & D. K. Harman (Eds.), *The Eighth Text REtrieval Conference (TREC-8).* Gaithersburg, MD, USA.

Larson, R. R. (2000). Berkeley's TREC-8 Interactive Track Entry: Cheshire and Zprise. In E. M. Voorhees & D. K. Harman (Eds.), *The Eighth Text REtrieval Conference (TREC-8).* Gaithersburg, MD, USA.

Yang, K., & Maglaughlin, K. L. (2000). IRIS at TREC-8. In E. M. Voorhees & D. K. Harman (Eds.), *The Eighth Text REtrieval Conference (TREC-8).* Gaithersburg, MD, USA.

The Proceedings of the Eighth Text REtrieval Conference will appear in print in mid-2000. Softcopies of most papers are available now from the TREC website (trec.nist.gov).