TREC-8 Interactive Track Report

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

October 25, 2000

Abstract

This report is an introduction to the work of the TREC-8 Interactive Track with its goal of investigating interactive information retrieval by examining the process as well as the results.

Seven research groups ran a total of 14 interactive information retrieval (IR) system variants on a shared problem: a question-answering task, six statements of information need, and a collection of 210,158 articles from the Financial Times of London 1991-1994.

This report summarizes the shared experimental framework, which for TREC-8 was designed to support analysis and comparison of system performance only within sites. The report refers the reader to separate discussions of the experiments performed by each participating group — their hypotheses, experimental systems, and results. The papers from each of the participating groups and the raw and evaluated results are available via the TREC home page (trec.nist.gov).

1 Introduction

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
- 6 topics — brief statements of information need
- a document collection to be searched
- 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 research groups

The framework allowed groups to estimate the effect of their experimental manipulation free and clear of the main (additive) effects of searcher and topic

Table 1: Participating research groups, their systems, and the number of searches performed on each.

Group	Searches
New Mexico State University at Las Cruces	72
Oregon Health Sciences University	144
Royal Melbourne Institute of Technology / CSIRO	144
Rutgers University	216
Sheffield University	144
University of California at Berkeley	72
University of North Carolina at Chapel Hill	144

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). Table 1 lists the research groups that took part and the total number of searches performed as part of their experiment. The issues addressed by each team are discussed in section 3.

2 Method

2.1 Participants

Each research group selected its own experimental participants, known here as "searchers." There was only one restriction: no searcher could have previously used either the control system or the experimental system. Additional restrictions were judged impractical given the difficulty of finding searchers. A minimum of twelve searchers was required, but the experimental design allowed for the addition of more in groups of four and additions were encouraged. Standard demographic data about each searcher were collected by each site and some sites administered additional tests.

2.2 Apparatus

IR systems

In addition to running its experimental system(s), each participating site chose a control system appropriate to the local research goals.

Computing resources

Each participating group was responsible for its own computing resources adequate to run both the control and experimental systems and collect the data required for its own experiments and for submission to NIST. 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.

Topics

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 as follows:

- 408i tropical storms
- 414i Cuba, sugar, imports
- 428i declining birth rates

- 431i robotic technology
- 438i tourism, increase
- 446i tourists, violence

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 allotted, 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.
```

Searcher task

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.

Table 2: Basic 2x2 Latin square on which evaluation is based.

Searchers	System,Topic combinations	
S1	E,Tx	C,Ty
S2	C,Ty	E,Tx

Document collection

The collection of documents to be searched was the Financial Times of London 1991-1994 collection (part of the TREC-8 adhoc collection). This collection contains 210,158 documents (articles) totaling 564 megabytes. The median number of terms per document is 316 and the mean is 412.7.

2.3 Procedure

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.

The minimal 12-searcher-by-6-topic matrix can be rearranged and seen as 18 2-searcher-by-2-topic Latin squares. Each 2-by-2 square has the form shown in Table 2 and has the property that the "treatment effect," here $E - C$, the control-adjusted response, can be estimated free and clear of the main (additive) effects of searcher and topic. Participant and topic are treated statistically as blocking factors. This means that even in the presence of the anticipated differ-

Table 3: Half the minimal 8-searcher-by-8-topic matrix as run.

Searchers	System,Topic combinations (in example order as seen by searchers)					
S1	E,T6	E,T1	E,T2	C,T3	C,T4	C,T5
S2	C,T1	C,T2	C,T3	E,T4	E,T5	E,T6
S3	C,T2	C,T3	C,T4	E,T5	E,T6	E,T1
S4	C,T3	C,T4	C,T5	E,T6	E,T1	E,T2
S5	E,T4	E,T5	E,T6	C,T1	C,T2	C,T3
S6	E,T5	E,T6	E,T1	C,T2	C,T3	C,T4
S7	C,T6	C,T1	C,T2	E,T3	E,T4	E,T5
S8	E,T1	E,T2	E,T3	C,T4	C,T5	C,T6
S9	E,T2	E,T3	E,T4	C,T5	C,T6	C,T1
S10	E,T3	E,T4	E,T5	C,T6	C,T1	C,T2
S11	C,T4	C,T5	C,T6	E,T1	E,T2	E,T3
S12	C,T5	C,T6	C,T1	E,T2	E,T3	E,T4

Table 4: Results by topic.

Topic	Mean instance recall across all searcher-systems	Mean instance precision across all searcher-systems	Number of instances identified by NIST
408i	0.326	0.777	24
414i	0.532	0.660	12
428i	0.306	0.667	26
431i	0.329	0.821	40
438i	0.172	0.734	56
446i	0.227	0.517	16

ences between searchers and topics, the designs provided estimates of $E - C$ that were not contaminated by these differences.

However, the estimate of $E - C$ would be contaminated by the presence of an interaction between topic and searcher. Therefore, we replicated the 2x2 Latin square 6x3 times to get the minimal 12x6 design for each site. The contaminating effect of the topic by searcher interaction was reduced by averaging the eighteen estimates of $E - C$ that are available, one for each 2x2 Latin square. This is analogous to averaging replicate measurements of a single quantity in order to reduce the measurement uncertainty. Each 2-by-2 square yields 1 within-searcher estimate of the $E - C$ difference for a total of 18 such estimates for each 12-searcher-by-6-topic matrix.

In resolving experimental design questions not covered here (e.g., scheduling of tutorials and searches, etc.), participating sites were asked to minimize the differences between the conditions under which a given searcher used the control and those under which he or she used the experimental system.

2.4 Data submitted to NIST

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).

- 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
- 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.

2.5 Evaluation of the sparse-format data submitted to NIST

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:

1. Read the topic carefully.
2. Read each of the documents from the pool for that topic and gradually:
   (a) Create a list of the instances found somewhere in the documents
   (b) Select and record a short phrase describing each instance found
   (c) Determine which documents contain which instances
   (d) Bracket each instance in the text of the document in which it was found

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.

3 Results and Discussion

The mean results by topic are presented here in Table 4. For TREC-8, topic presentation sequence was randomized for each searcher.

A summary of each group's results (instance recall by site and condition) is shown in Table 5. Comparison of systems across sites is not supported by the experimental design, so comparisons presented here are between systems within a given site. A general theme running through the results was that there was little difference between each group's experimental and control systems. Whether it was New Mexico State University's document summarization approach or the use of the relevance feedback by Sheffield University, users showed little difference across systems, many of which contained features shown to be effective in non-interactive experiments in the past. A generalization can be made that 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.

Table 5: Instance recall by site and condition.

Site	Condition	Instance recall
NMSU	Added document summaries	0.44
	Baseline full text	0.40
OHSU	Okapi weighting	0.38
	Baseline tf*idf weighting	0.33
RMIT/CSIRO	Added categorization interface	0.27
	Baseline document list interface	0.31
Rutgers	Relevance feedback	0.26
	Local context analysis	0.24
Sheffield	With relevance feedback	0.35
	Without relevance feedback	0.39
Berkeley	Enhanced Cheshire interface	0.38
	Baseline ZPRISE Interface	0.41
UNC	Passage-level retrieval feedback	0.23
	Document-level retrieval feedback	0.28

There are two possible explanations for this. Either there really is no difference or these experiments lack the research design or statistical power to detect a difference. Only further research, including the study other types of search tasks and larger numbers of queries, will resolve this.

The actual results obtained by each group are summarized in the following paragraphs. For more details the reader is directed to the site reports in these proceedings or on the TREC web site (trec.nist.gov).

- 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 in these proceedings (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 in these proceedings (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 in these proceedings (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 in these proceedings (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 in these proceedings (Larson, 2000).

- University of North Carolina found no difference among various levels of relevance feedback. For more information see the site report in these proceedings (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 paramaterization 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).

4 Authors' note

The design of the TREC-8 Interactive Track matrix experiment grew out of the efforts of the many people who contributed to the discussion of ends and means on the track discussion list and through other channels.

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.

5 Appendix: Instructions to be given to each searcher

The following introductory instructions are to be given once to each searcher before the first search:

> Imagine that you have just returned from a visit to your doctor during which it was discovered that you are suffering from high blood pressure. The doctor suggests that you take a new experimental drug, but you wonder what alternative treatments are currently available. You decide to investigate the literature on your own to satisfy your need for information about what different alternatives are available to you for high blood pressure treatment. You really need only one document for each of the different treatments for high blood pressure.

You find and save a single document that lists four treatment drugs. Then you find and save another two documents that each discusses a separate alternative treatment: one that discusses the use of calcium and one that talks about regular exercise. You've run out of time and stop your search. In all, you have identified six different instances of alternative treatments in three documents.

In this experiment, you will face a similar task. You will be presented with several descriptions of needed information on a number of topics. In each case there can be multiple examples or instances of the type of information that's needed.

We would like you to identify as many different instances as you can of the needed information for each topic that will be presented to you - as many as you can in the 20 minutes you will be given to search. Please save one document for EACH DIFFERENT instance of the needed information that you identify. If you save one document that contains several instances, try not to save additional documents that contain ONLY those instances. However, you will not be penalized if you save documents unnecessarily.

As you identify an instance of the needed information, please keep track of which instances you have found: write down a word or short phrase to identify the instance, or–if the system provides a facility to keep track of instances–use it.

Carefully read each topic to understand the type of information needed. This will vary from topic to topic. On one topic you may be looking for instances of a certain kind of event. On another you may be searching for examples of certain sorts of people, places, or things.

Do you have any questions about

- what we mean by instances of needed information,
- the way in which you are to save nonredundant documents for each instance?