TREC-9 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
Retrieval Group
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA

May 14, 2001

Abstract

The TREC Interactive Track has the goal of investigating interactive information retrieval by examining the process as well as the results. In TREC-9 six research groups ran a total of 12 interactive information retrieval (IR) system variants on a shared problem: a fact-finding task, eight questions, and newspaper/newswire documents from the TREC collections. This report summarizes the shared experimental framework, which for TREC-9 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-9 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. In particular, the track examined the use of IR systems in a fact-finding task — searchers had to find the answers to questions designed to require reference to multiple documents. There was a strong desire to reduce the time per search (previously: 20 minutes), to reduce the overall search session time per searcher (more than three hours), to use different data from that used for the last several years by the track (the Financial Times of London), and to explore different types of questions from the sort studied in the last several TREC interactive tracks, ones which would require some simple organization of the found information. In response to these goals a common experimental framework was designed with the following features:

- an interactive search task — question answering
- 8 questions — short answers

- 16 searchers — minimum
- a newswire/newspaper article collection to be searched
- a required set of searcher questionnaires
- 5 classes of data to be collected at each site and submitted to NIST

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

In TREC-9 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 obtained. 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).

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 sixteen searchers was required, but the experimental design allowed for the addition of more in groups of eight 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.

Questions

Questions from the non-interactive TREC-8 Question and Answer Track were considered for use, but proved too easy for an interactive task. A number of candidate questions were developed by the participating research groups inspired more by the data at hand than any systematic considerations. Four sorts were considered and tested to gauge their suitability:

- Find any n Xs, e.g., Name three US Senators on committees regulating the nuclear industry.
- Comparison of two specific Xs, e.g., Do more people graduate with an MBA from Harvard Business School or MIT Sloan?
- Find the largest/latest/... n Xs, e.g., What is the largest expenditure on a defense item by South Korea?
- Find the first or last X, e.g., Who was the last Republican to pull out of the nomination race to be the candidate of his/her party for US president in 1992?

In the end, eight questions were chosen, four of each of the first two types. Questions of the last two sorts were difficult to find/create and, given their "superlative" nature, seemed less likely to be doable in the five minutes alloted to each search. All the questions called for very short answers.The first four required the searcher to respond with an answer that has from one to four parts. This was a bounded version of the instance retrieval type of question used in TREC-5 through TREC-8 interactive tracks. The second four required the searcher to decide which of two given answers is the correct one. Here are the questions:

1. What are the names of three US national parks where one can find redwoods?
2. Identify a site with Roman ruins in present day France.
3. Name four films in which Orson Welles appeared.
4. Name three countries that imported Cuban sugar during the period of time covered by the document collection.
5. Which children's TV program was on the air longer: the original Mickey Mouse Club or the original Howdy Doody Show?
6. Which painting did Edvard Munch complete first: "Vampire" or "Puberty"?
7. Which was the last dynasty of China: Qing or Ming?
8. Is Denmark larger or smaller in population than Norway?

Searcher task

The task of the interactive searcher was to find and record the answer to the question and identify one or more documents that supported the answer — all within the five minutes allotted for each question. The question creation process guaranteed that each question could be answered based on documents in the collection.

Document collection

The collection of documents to be searched included the following TREC collections:

1. Associated Press (disks 1-3)
2. Wall Street Journal (disks 1-2)
3. San Jose Mercury News (disk 3)
4. Financial Times from (disk 4)
5. Los Angeles Times (disk 5)
6. Foreign Broadcast Information Service (disk 5)

Table 1: Minimal 16-searcher-by-8-question matrix as run.

Searcher	Block 1 System: Questions	Block 2 System: Questions
1	B: 4-7-5-8	A: 1-3-2-6
2	A: 3-5-7-1	B: 8-4-6-2
3	A: 1-3-4-6	B: 2-8-7-5
4	A: 5-2-6-3	B: 4-7-1-8
5	B: 7-6-2-4	A: 3-5-8-1
6	B: 8-4-3-2	A: 6-1-5-7
7	A: 6-1-8-7	B: 5-2-4-3
8	B: 2-8-1-5	A: 7-6-3-4
9	A: 4-7-5-8	B: 1-3-2-6
10	B: 3-5-7-1	A: 8-4-6-2
11	B: 1-3-4-6	A: 2-8-7-5
12	B: 5-2-6-3	A: 4-7-1-8
13	A: 7-6-2-4	B: 3-5-8-1
14	A: 8-4-3-2	B: 6-1-5-7
15	B: 6-1-8-7	A: 5-2-4-3
16	A: 2-8-1-5	B: 7-6-3-4

2.3 Procedure

Each searcher performed eight searches on the document collection using the eight interactive track topics in a pseudo-random order. Each searcher performed 4 searches on one of the site's systems and then 4 on the other to avoid the extra cognitive load of switching systems with each search. Table 1 shows an example ordering of searches for two systems, eight questions, and sixteen searchers. 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.

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

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

The minimal 16-searcher-by-8-topic matrix can be rearranged and seen as 32 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 differences between searchers and topics, the design 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 2-by-2 Latin square 8x4 times to get the minimal 16-by-8 design for each site. The contaminating effect of the topic by searcher interaction was reduced by averaging the thirty-two estimates of $E - C$ that are available, one for each 2-by-2 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 one within-searcher estimate of the $E - C$ difference for a total of thirty-two such estimates for each 16-searcher-by-8-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

Five sorts of data were collected for evaluation/analysis (for all searches unless otherwise specified) and are available from the TREC-9 Interactive Track web page (www-nlpir.nist.gov/projects/t9i).

- 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 satisfaction, etc.
- a full narrative description of one interactive session for a question to be chosen by each site
- any further guidance or refinement of the task specification given to the searchers

Only the sparse-format data were evaluated at NIST. Each each response, i.e., each attempt to answer a question, was assessed using two questions:

- Does the response contains all, some or none of the items asked for by the question?
- Do the documents cited fully support all, some or none of the correct items in the response?

Note that in the case of the "Is it A or B" questions (5 - 8), the response can contain at most one item, so the answer to the first assessment question can only be "all" items (i.e., one), or "none", and similarly for the second question. Counts for partial responses and partial support are thus not present in the next section's assessment outcome figures for this sort of question.

3 Results and Discussion

This section presents the raw results aggregated across all sites and systems by question. The total number of responses per question varies since not all sites submitted complete results. Each table presents the number of responses in each assessment category. All the questions of type "Is it A or B?" are presented first with their reduced set of possible outcomes; otherwise the tables are presented in order of decreasing success. Discussion of the supporting documents is limited to those submitted with the responses; no exhaustive search of the document collection was undertaken to find all possible supporting documents.

3.1 Questions

Question 7 had 23 different documents submitted in support of answers to it (see Figure 1.) and the assessors found all to be supportive. Twelve of the documents provide the dates for the Qing dynasty only, seven for the Ming only, and four the combined dates without allowing one to say which came first.

Question 5 had only seven documents saved in support of it (see Figure 2.) and all were supportive to some extent. Four provided the dates for the Howdy Doody Show and three for the Mickey Mouse Club. So, one document of each sort was needed for a fully supported answer.

Question 6 had only two supportive documents (see Figure 3.) — one for "Vampire" and one for "Puberty". Seven documents were submitted as supportive.

Question 8 (see Figure 4.) was answered in part by ten documents that provided the population of Denmark and five that included the number for Norway. There were no documents that included both numbers.

Question 4 (see Figure 5.) had 54 documents submitted as supportive but the assessors found only 39 to be so. Possible answers (with the number of documents providing them in parentheses were Indonesia (23), Khazakhstan (19), South Korea (19), former Soviet Union (3), Soviet Union (16), Russia (10), China (5), Canada (3), Japan (3), Latvia (2), Britain/UK (1), Caricom (1), Eastern Europe / E. Germany (1), Iran (1), Italy (1), Mexico (1), Portugal (1), and Socialist Bloc (1). Seven percent of all responses contained no answer. Twenty percent of all responses contained an incorrect answer (15% contained 1 wrong answer, 4% contained 2, 1% contained 3).

Figure 1: Responses to question 7 by assessment outcome.

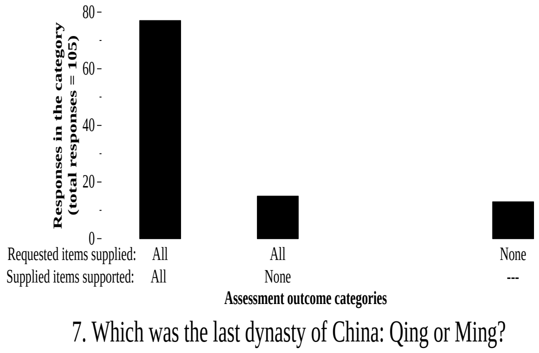


Figure 2: Responses to question 5 by assessment outcome.

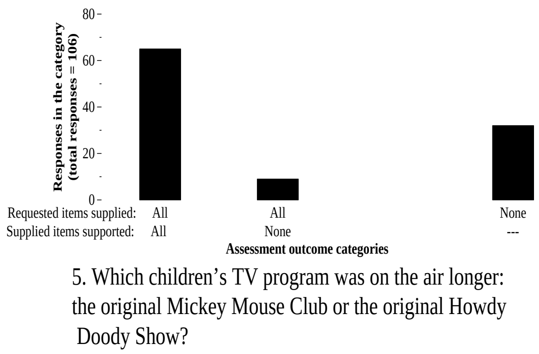


Figure 3: Responses to question 6 by assessment outcome.

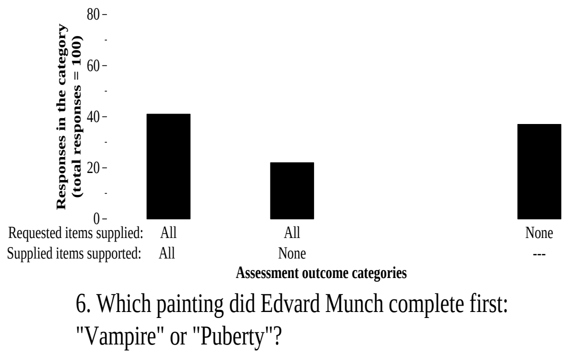
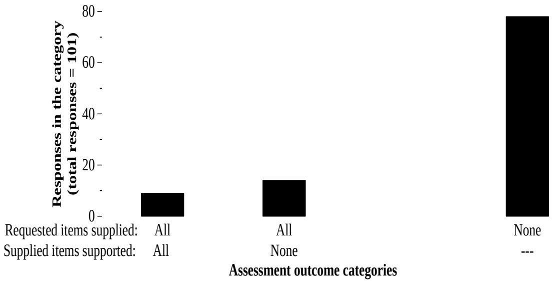
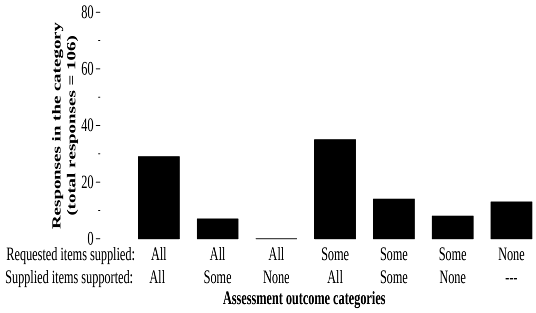


Figure 4: Responses to question 8 by assessment outcome.



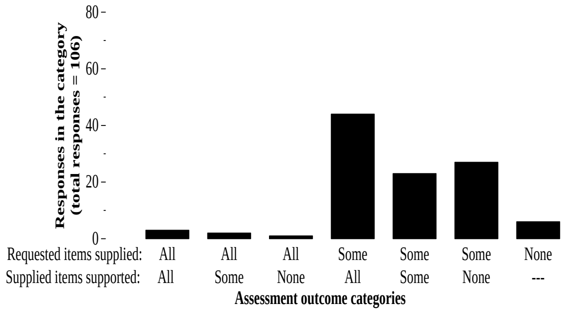
8. Is Denmark larger or smaller in population than Norway?

Figure 5: Responses to question 4 by assessment outcome.



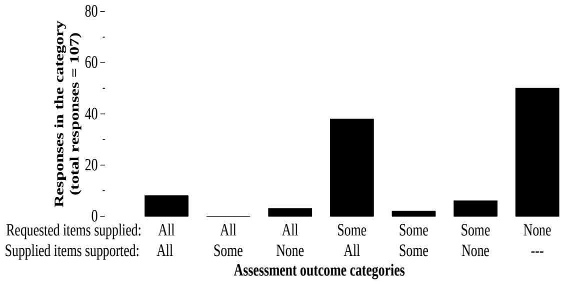
4. Name three countries that imported Cuban sugar....

Figure 6: Responses to question 3 by assessment outcome.



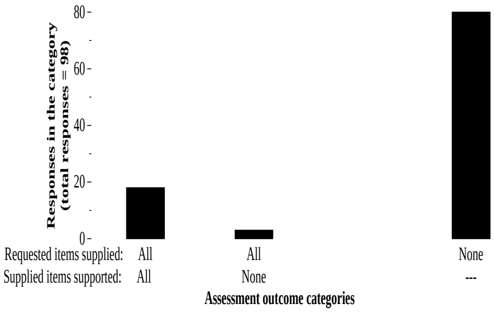
3. Name four films in which Orson Welles appeared.

Figure 7: Responses to question 1 by assessment outcome.



1. What are the names of 3 US national parks where one can find redwoods?

Figure 8: Responses to question 2 by assessment outcome.



2. Identify a site with Roman ruins in present day France.

For question 3 (see Figure 6.), 40 documents were saved as supportive but only 17 were judged to be so. Possible answers were Citizen Kane (4), Third Man (4), Catch-22 (2), Othello (2), Some to Love (2), Chimes at Midnight (1), Lady from Shanghai (1), and MacBeth (1). The collection contained a number of references to films which Welles directed but did not appear in. In the haste of the moment, some searchers may have overlooked this important distinction. Five percent of all responses contained no answer. Forty-nine percent of all responses contained an incorrect answer (33% contained 1 wrong answer, 9% contained 2, 7% contained 3). "The Magnificent Ambersons" accounts for 80% of the responses with a single wrong answer. This film was directed by Welles and his was the voice of the narrator. The assessor did not consider this an appearance.

For question 1 (see Figure 7.), 24 documents were submitted, of which only 13 were supportive. Possible answers were Redwood National Park (5), Sequoia National Park (5), Yosemite National Park (5), Kings Canyon National Park (4), California Six Rivers National Park (1), and Lassen National Park (4). The collection contained many references to state parks with redwoods and some of these were submitted as answers but were not counted as valid. There may also have been some question about whether a sequoia is redwood and whether a national monument or national forest should count as a national park. The assessors answered "yes" to all those questions. Fifteen percent of all responses contained no answer. Forty-two percent of all responses contained an incorrect answer (9% contained 1 wrong answer, 8% contained 2, 24% contained 3).

Finally, for question 2 (see Figure 8.), 27 documents were submitted as supportive but only 7 were found to be so. Possible answers found were amphitheater in southern France (2), arena at Nimes (2), ruins in Arles (2), arena of Lutec (1), ruins in Orange (1), ruins near Frethun (1), and ruins near Perigord, North Dordogue (1). Forty-eight percent of all responses contained no answer. Thirty-four percent of all responses contained an incorrect answer. The question required only one item per answer.

3.2 Approaches

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

- Chapman University (Vogt, in press)investigated the use of a rich transcript of user actions to predict relevance of documents viewed.
- Glasgow University (Alexander, Brown, & Joemon, in press) looked at the value of summaries:
  - indicative, query-biased document summaries
  - full text of documents
- Oregon Health Sciences University (Hersh et al., in press) asked whether techniques which are effective in batch IR are also effective in an interactive setting. Their work compared Okapi weighting with tf.idf weighting.
- Royal Melbourne Institute of Technology-CSIRO (D'Souza, Fuller, Thom, Vines, & Zobel, in press)compared the use of two different document surrogates:
  - document title plus the first twenty words from the document
  - document title plus the three "best" sentences

  They used two measures of system effectiveness: number of responses complete and fully supported and number of requested items correct and fully supported.
- Rutgers University (Belkin et al., in press) examined two interfaces for question answering:
  - 10 titles plus the text of the top document plus suggested terms
  - 6 scrollable documents showing the "best passage"

They evaluated the systems in terms of number of responses complete and fully supported.

- Sheffield University (Beaulieu, Fowkes, & Joho, in press) studied a known system's (Okapi) performance on the new task.

3.3 Future work

Results from the TREC Interactive Track have shown over the last few years that interactive evaluation, while complicated, is possible and can generate informative results. There is agreement among most of the track participants, however, that there is room for methodological improvements within the basic TREC setting. A workshop was held at SIGIR 2000 to explore such possible improvements (Hersh & Over, 2000). The recommendations are listed below. They will be the basis for the design of the TREC-2001 Interactive Track, which will comprise focused observational studies of Web searching. It is hoped that from the observations will come the germs of hypotheses which can be implemented and tested in a more controlled experimental setting for TREC-2002.

The SIGIR workshop's recommendations are as follows:

- Relieve some of the pressure on participants by running the track on a 2-yr cycle with interim results reported after the first 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, etc.

- Define Web search tasks in four domains chosen based on surveys of popular web usage - tasks experimental searchers should be able to identify with based on a simple cover story: finding consumer 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.

- At least for the first 2-yr cycle (TREC-2001/2) allow participants to undertake mainly observational studies during the first year, but designed to support metrics-based comparison of systems during the second year. This might involve collecting web documents during year 1 for use as a static collection in year 2.

- Alter the experimental design (probably only for use in year 2) to allow for more statements of information need e.g., questions (circa 25). A given searcher would only search a small subset. It might still be based on the 2-topic-by-2-search Latin square to retain blocking by searcher and topic.

4 Authors' note

The design of the TREC-9 Interactive Track matrix experiment grew out of the efforts many people, who contributed to the discussion the track discussion list, suggested questions, and helped test them.

References

Alexander, N., Brown, C., & Joemon, J. (in press). Question answering, relevance feedback and summarisation: TREC-9 interactive track report. In E. M. Voorhees & D. K. Harman (Eds.), *The Ninth Text REtrieval Conference (TREC-9).* Gaithersburg, MD, USA.

Beaulieu, M., Fowkes, H., & Joho, H. (in press). Sheffield Interactive Experiment at TREC-9. In E. M. Voorhees & D. K. Harman (Eds.), *The Ninth Text REtrieval Conference (TREC-9).* Gaithersburg, MD, USA.

Belkin, N. J., Keller, A., Kelly, D., Perez-Carballo, J., Sikora, C., & Sun, Y. (in press). Support for Question-Answering in Interactive Information Retrieval: Rutgers' TREC-9 Interactive Track Experiments. In E. M. Voorhees & D. K. Harman (Eds.), *The Ninth Text REtrieval Conference (TREC-9).* Gaithersburg, MD, USA.

D'Souza, D., Fuller, M., Thom, J., Vines, P., & Zobel, J. (in press). Melbourne TREC-9 Experiments. In E. M. Voorhees & D. K. Harman (Eds.), *The Ninth Text REtrieval Conference (TREC-9).* Gaithersburg, MD, USA.

Hersh, W., & Over, P. (2000). SIGIR Workshop on Interactive Retrieval at TREC and Beyond. *SIGIR Forum*, *34*(1), 24–27.

Hersh, W., Turpin, A., Sacherek, L., Olson, D., Price, S., Chan, B., & Kraemer, D. (in press). Further Analysis of Whether Batch and User Evaluations Give the Same Results With a Question-Answering Task. In E. M. Voorhees & D. K. Harman (Eds.), *The Ninth Text REtrieval Conference (TREC-9).* Gaithersburg, MD, USA.

Vogt, C. C. (in press). Passive Feedback Collection – An Attempt to Debunk the Myth of Clickthroughs. In E. M. Voorhees & D. K. Harman (Eds.), *The Ninth Text REtrieval Conference (TREC-9).* Gaithersburg, MD, USA.