TREC-7 Interactive Track Report

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

July 12, 1999

Abstract

This report is an introduction to the work of the TREC-7 Interactive Track with its goal of investigating interactive information retrieval by examining the process as well as the results.

Eight research groups ran a total of 15 interactive information retrieval (IR) systems on a shared problem: a question-answering task, eight 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-7 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-7 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
- 8 topics - brief statements of information need
- a document collection to be searched
- a required set of searcher questionnaires
- a required psychometric test for all searchers
- 6 classes of data to be collected at each site and submitted to NIST
- 3 summary measures to be calculated by NIST for use by participants

The framework allowed groups to estimate the effect of their experimental manipulation free and clear of the main (additive) effects of participant and topic and it was designed to reduce the effect of interactions.

In TREC-7 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). Figure 1 lists the research groups that took part, their systems (control and experimental), and the number of searches performed on each. Here are the high-level issues addressed by each team:

- The researchers at New Mexico State University at Las Cruces investigated the benefit of a thumbnail-document view over a more conventional interface. (Ogden, Davis, & Rice, 1999)
- In London and Sheffield the Okapi Group made two pairwise comparisons: Okapi with relevance feedback versus Okapi without and Okapi without versus ZPRISE without. (Robertson, Walker, & Beaulieu, 1999)
- The team at Oregon Health Sciences University carried out a large-scale comparison of Boolean and natural language searching involving 28 searchers. (Hersh et al., 1999)
- The Royal Melbourne Institute of Technology group examined differences in retrieval coverage and efficiency resulting from different organizations of query results: a list of cluster descriptors versus a list of document titles. (Fuller et al., 1999)
- Researchers at Rutgers conducted a study to investigate the effectiveness and usability of a particular implementation of negative relevance feedback and of relevance feedback as a term-suggestion device. (Belkin et al., 1999)
- At the University of California at Berkeley they replicated their experiments from TREC-6 but with larger numbers of searchers/searches and more information about searchers and their search experiences gathered from the track questionnaires. (Gey, Jiang, Chen, & Larson, 1999)
- The University of North Carolina at Chapel Hill team tried to determine whether the ability to see and modify term weights improves retrieval effectiveness and whether the passage is a better unit of relevance feedback than the document. (Yang, Maglaughlin, Meho, & Sumner, 1999)
- The University of Toronto researchers compared an experimental system which blended querying and browsing with a system approximating a common web search engine system where querying is distinct from browsing the documents found. (Bodner & Chignell, 1999)

Groups	Systems	Searches
New Mexico State University at Las Cruces	J24	32
	ZP	32
Okapi Group	ok_noRF	32
	zp_noRF	32
	ok_noRF	32
	ok_withRF	32
Oregon Health Sciences University	MB	112
	MR	112
Royal Melbourne Institute of Technology	clus	64
	list	64
Rutgers University	RUINQ-G	67
	RUINQ-R	68
University of California at Berkeley	C	32
	Z	32
University of North Carolina at Chapel Hill	irisp	32
	iriss	32
	irisa	32
	iriss	32
University of Toronto	a	32
	b	32

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

2 Method

2.1 Participants

Each research group selected its own experimental participants, known in what follows 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 eight 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 appro-

priate 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 their 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

Eight of the 50 topics created by NIST for the TREC-7 adhoc task were selected and modified for use in the interactive track by adding a section called "Instances" and removing the "Narrative." The eight topics were entitled as follows:

- 352i British Chunnel impacts
- 353i antarctic exploration
- 357i territorial waters dispute
- 362i human smuggling
- 365i El Nino
- 366i commerical cyanide uses
- 387i radioactive waste
- 392i robotics

Each of the eight 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: 352i

Title:  British Chunnel impacts

Description:
Impacts of the Chunnel - anticipated
or actual - on the British economy
and/or the life style of the British

Instances:
In the time alloted, please find as
many DIFFERENT impacts of the sort
described above as you can. Please
save at least one document for EACH
such  DIFFERENT impact. If one
document discusses several such
impacts, then you need not save
other documents that repeat those,
since your goal is to identify as
many DIFFERENT impacts of the sort
described above as possible.
```

The results of test searches performed at NIST were used to:

- choose the eight topics from a larger set
- attempt to balance the blocks for difficulty
- attempt to define the sequence of use within each block so that the difficulty increased

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 15 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.

Document collection

The collection of documents to be searched was the Financial Times of London 1991-1994 collection (part of the TREC-7 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.

Searchers	System,Topic combinations (in Latin squares for evaluation)							
S1	E,T1	C,T5	E,T2	C,T6	E,T3	C,T7	E,T4	C,T8
S2	C,T5	E,T1	C,T6	E,T2	C,T7	E,T3	C,T8	E,T4
S3	E,T5	C,T1	E,T6	C,T2	E,T7	C,T3	E,T8	C,T4
S4	C,T1	E,T5	C,T2	E,T6	C,T3	E,T7	C,T4	E,T8

Figure 2: Half the minimal 8-searcher-by-8-topic matrix as evaluated. E = experimental system, C = control.

Searchers	System,Topic combinations (in the order seen by searchers)							
S1	E,T1	E,T2	E,T3	E,T4	C,T5	C,T6	C,T7	C,T8
S2	C,T5	C,T6	C,T7	C,T8	E,T1	E,T2	E,T3	E,T4
S3	E,T5	E,T6	E,T7	E,T8	C,T1	C,T2	C,T3	C,T4
S4	C,T1	C,T2	C,T3	C,T4	E,T5	E,T6	E,T7	E,T8

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

2.3 Procedure

Each searcher performed eight searches on the document collection using the eight interactive track topics. Each searcher performed half of the total number of searches on the site's experimental system and the other half on its control system. 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 order in which each searcher used the systems (experimental and control) and topics.

The minimal 8-searcher-by-8-topic matrix was constructed of 16 2-searcher-by-2-topic Latin squares. Figure 2 shows half of the required matrix; the other half is identical except it includes four additional searchers. Each 2-by-2 square 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 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 4x4 times to get the minimal 8x8 design for each site. The contaminating effect of the topic by searcher interaction was reduced by averaging the sixteen 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 16 such estimates for each 8-searcher-by-8-topic matrix.

To reduce the searcher's cognitive load and possible confusion due to switching search systems with each search, the columns were permuted as indicated in Figure 3 for the running of the experiment.

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 other-

wise specified) and are available from the TREC-7 Interactive Track web page (www-nlpir.nist.gov/projects/t7i/t7i.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.
- the results of the Educational Testing Service's FA-1 test (controlled associations)
- a full narrative description of one interactive session for topic 365i
- 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

Since comparison of systems across sites is not supported by the experimental design, the reader is directed to the site reports in these proceedings or on the TREC web site (trec.nist.gov) for presentation and discussion of results in context of the local research goals.

The mean results by topic are presented here in Figure 4. Since thh order of topics was the same in all experiments, the effect of order is indistinguishable from that of topic. While the first topic in each block seems to have been easier than those that followed, it is not clear that the blocks are overall of equal difficulty, as was intended.

4 Author's note

The design of the TREC-7 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. The author would like to acknowledge the special contributions of the track coordinators, Steve Robertson and Nick Belkin, of Bill Hersh, who coordinated the use of the FA-1 test.

Block	Order in block	Topic	Mean instance recall across all searcher-systems	Mean instance precision across all searcher-systems	Number of searches	Number of instances identified by NIST
1	1	365i	0.750	0.893	117	24
	2	357i	0.257	0.437	117	13
	3	362i	0.259	0.632	117	12
	4	352i	0.248	0.673	117	28
2	1	366i	0.375	0.835	117	7
	2	392i	0.324	0.692	117	36
	3	387i	0.375	0.778	117	9
	4	353i	0.187	0.409	116	11

Figure 4: Results by topic.

References

Belkin, N. J., Perez Carballo, J., Cool, C., Kelly, D., Lin, S., Park, S. Y., Rieh, S. Y., Svage-Knepshield, P., & Sokora, C. (1999). Rutgers' TREC-7 Interactive Track Experience. In E. M. Voorhees & D. K. Harman (Eds.), *The Seventh Text REtrieval Conference (TREC-7).* Gaithersburg, MD, USA.

Bodner, R. C., & Chignell, M. H. (1999). ClickIR: Text Retrieval using a Dynamic Hypertext Interface. In E. M. Voorhees & D. K. Harman (Eds.), *The Seventh Text REtrieval Conference (TREC-7).* Gaithersburg, MD, USA.

Fuller, M., Kaszkiel, M., Kim, D., Ng, C., Robertson, J., Wilkinson, R., Wu, M., & Zobel, J. (1999). TREC 7 Ad Hoc, Speech, and Interactive Tracks. In E. M. Voorhees & D. K. Harman (Eds.), *The Seventh Text REtrieval Conference (TREC-7).* Gaithersburg, MD, USA.

Gey, F., Jiang, H., Chen, A., & Larson, R. R. (1999). Manual Queries and Machine Translation in Cross-language Retrieval and Interactive Retrieval with Cheshire II at TREC-7. In E. M. Voorhees & D. K. Harman (Eds.), *The Seventh Text REtrieval Conference (TREC-7).* Gaithersburg, MD, USA.

Hersh, W., Price, S., Kraemer, D., Chan, B., Sacherek, L., & Olson, D. (1999). A large-scale comparison of boolean vs. natural-language searching for the TREC-7 interactive track. In E. M. Voorhees & D. K. Harman (Eds.), *The Seventh Text REtrieval Conference (TREC-7).* Gaithersburg, MD, USA.

Ogden, W., Davis, M., & Rice, S. (1999). Document thumbnail visualizations for rapid relevance judgements: When do they pay off? In E. M. Voorhees & D. K. Harman (Eds.), *The Seventh Text REtrieval Conference (TREC-7).* Gaithersburg, MD, USA.

Robertson, S. E., Walker, S., & Beaulieu, M. (1999). Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. In E. M. Voorhees & D. K. Harman (Eds.), *The Seventh Text REtrieval Conference (TREC-7).* Gaithersburg, MD, USA.

Yang, K., Maglaughlin, K., Meho, L., & Sumner, R. G., Jr. (1999). IRIS at TREC-7. In E. M. Voorhees & D. K. Harman (Eds.), *The Seventh Text REtrieval Conference (TREC-7).* 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 15 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?