TREC 2002 Interactive Track Report

William Hersh, Oregon Health & Science University, Portland, OR, USA

For TREC 2002, the Interactive Track completed the two-year cycle of observational studies begun in TREC 2001 and followed now by more controlled laboratory experiments focusing on question answering using Web data. Six research groups participated this year.

Background

The Interactive Track was one of the first track's at TREC (dating back to TREC-3) and has always had a small but dedicated following. The track has accomplished much during its existence, including a special issue of *Information Processing and Management* (May/June, 2001) [1]. A variety of findings have emerged from experiments carried out by track participants:

- Presentation of documents matters to users, i.e., better surrogates help but clustering does not [2]
- Users do not utilize relevance feedback as we might think/hope [3]
- Results from "batch" studies do not necessarily apply to real-world searchers [4]

However, there have been limitations to the generalizability of the results obtained:

- Study sample sizes are small and conditions are artificial
- Searcher populations used might not be generalizable
- But the same apply to non-interactive studies, e.g.,
    - Is one search on many queries any better than multiple searches on the same query?
    - Some of the queries from batch experiments given to real users showed they were too easy

To help set the future direction of the track, a workshop was held at SIGIR 2000 [5]. It was decided at the workshop and subsequently that the track would move to a two-year cycle that would allow increased data collection to better formulate and study hypotheses. In particular, in the first year of the cycle, groups would perform observational studies that increased the realism of task and generated experimental hypotheses for the following year. The TREC 2001 Interactive Track was first year of observational studies, with hypothesis-driven experiments to be performed in TREC 2002.

Data for searching

The track used on open version of the .GOV Web collection created for the TREC 2002 Web Track for searching. The collection was "open" in the sense that some links to pages outside the collection were presented and could be followed. This meant that the collection was intermediate in its stability between the live Web used by the track last year and a completely fixed, closed version of the .GOV collection, which was desired but not available in time for experimentation.

The collection was used by most participating groups as indexed and searched by the Panoptic search engine. The cited version does stemming and the homepage-finding feature is turned off. Results could be obtained in XML format by sending a query via CGI, e.g.,

trec.panopticsearch.com/gov/padre-sw_xml.cgi?collection=gov&query=bush

and getting back a padre_results packet. Experiments did not need to be limited to the interface defined by actual HTML pages returned by the Panoptic engine. Help on the use of the Panoptic search engine was available.

Tasks

Eight searcher tasks, analogous to those used in TREC 2001, were generated and tested. They are listed below. All searchers were to be given at least 10 minutes on each task once the actual searching began, and participating groups had to report the results as of the end of the ten-minute period in their proceedings papers. Groups optionally were able to report additional results, e.g., after 5 minutes, 15 minutes, etc..

There were four general searching activities from upon which the eight actual tasks were modeled. The four activities were:

1. Looking for personal health information
2. Seeking guidance on US government laws, regulations, guidelines, policy
3. Making travel plans
4. Gathering material for a report on a given subject

The searcher tasks were formulated in one of the following ways:

1. Find any N short answers to a question, to which there are multiple answers of the same type.
2. Find any N websites that meet the need specified in the task statement

The eight tasks proper were:

1. You are traveling from the Netherlands, and want to bring some typical food products as gifts for your friends. What are three kinds of food products from the Netherlands that you are not allowed to bring into the US? [Government Regulation]
2. You are concerned with privacy issues related to electronic information and would like to know what laws have been passed by the US Congress regarding these issues. Identify three such laws. [Government Regulation]
3. A friend has a private well which is the family's only source of drinking water. Locate a US publication, which contains guidelines for the maintenance of safe water standards for private well use. [Health]
4. You are not sure about the safety of genetically engineered foods, and would like to find more information and research on this topic. Name four potential types of safety problems that have been raised. [Health or Project]
5. You are interested in learning more about what measures the US government has taken since 2001 to prevent Mad-Cow Disease. Identify three such measures. [Health or Project]
6. Name/find three research programs/projects that investigate the treatment/causes of dwarfism. [Project]
7. You are planning a cycling expedition along the Silk Road in Central Asia. Find a website that is a good source information about health precautions should you take. [Travel]

8. You are planning to travel to the northeast territories of India and wonder if there are any problems/restrictions for tourists. Find a website that is a good source of information about such problems/restrictions. [Travel]

The various sites performed their own "grading" of the tasks and determination of relevant documents. An informal effort was made among groups to share answers and relevant documents.

A standard query was run by all searchers and the time between pressing the search button and the return of the results logged. This information was used to get an idea of the different response time conditions searchers may have encountered.

Experimental design

The experimental design followed the protocol developed for the TREC-9 Interactive Track. This design allowed the comparison of two systems or system variants. A minimum of 16 searchers were to be used. Each searcher was to perform all eight tasks (two of each of the four types), half on one system and half on another.

Each searcher was to issue the query "information retrieval" twice - once before beginning each set of the four searches on the two search engines in the design. The searcher ignored the results of the search. Experimenters collected elapsed time for these searches from the time the search button was pressed until results started to appear. This calibrating information was to be reported in each group's proceedings paper.

Evaluation

The searches were evaluated for effectiveness, efficiency, and user satisfaction in a manner similar to previous interactive tracks. Effectiveness included at least whether the task was completed successfully. Efficiency included at least the elapsed time used for each search. Instruments for the collection of minimal searcher background and satisfaction information were available and their use was encouraged.

Approaches

Here is a high-level description of the approaches taken by each group. For more detail, see the site report for each group.

CSIRO

In this year's Interactive Track, CSIRO continued to focus on answer organization issues, aiming to investigate whether the knowledge of "organizational structure" could be useful in organizing and delivering the retrieved documents. Particularly, for the collection of documents from the.gov domain, they used the level two domain name to categories the retrieved documents. For example, all documents from the nih.gov domain were put into the "National Institutes of Health" category. They compared this delivery method with the traditional ranked list. Their

preliminary results indicated that there was no significant difference between the two delivery methods in the first 5 minutes, but the subjects performed significantly better with the category interface at the end of 15 minutes.

Oregon Health & Science University

The Oregon group initially planned to compare retrieval using alternative devices, e.g., a tablet device, but was not able to do so when the vendor was unable to deliver the devices in time. Instead, they chose to revisit the search for factors associated with successful searching. Their results identified some trends but the sample size was inadequate in size to achieve statistical significance.

Rutgers University

The Rutgers group investigated two major hypotheses: (a) that reducing the amount of interaction required of a searcher with the system leads to increased satisfaction with the search and increased performance, and (b) that increased query length leads to increased performance in the TREC 2002 Interactive Track task. Both of these hypotheses were the result of their investigations in the TREC 2001 Interactive Track.

They investigated the first hypothesis by implementing two different interfaces to the Panoptic search engine: one which displayed the default Panoptic result of 20 URLs and snippets at a time, requiring the searcher to follow links in order to view pages, and a second which displayed, in scrollable windows, four retrieved pages at a time. The latter was intended to reduce interaction effort in comparison to the former, by virtue of displaying retrieved pages directly. They investigated the second hypothesis by having two different versions of each of the two interfaces: one which labeled the query input box, "QUERY," and another which labeled it, "Information problem (the more you say, the better the search results are likely to be)." The demonstration of the first version of query elicitation used only words and phrases; the demonstration of the second version used full sentences and questions.

University of North Carolina Chapel Hill

The North Carolina group's research question was whether 3D visualization of search results was more effective than a text-based interface. Unlike most of the other interactive track groups, they used their our own locally developed software. The 3D visualization was a variation on an Information Space navigation system (as they have demonstrated at past TREC conferences); the text system was a fairly generic Google-style interface, not much different from the Panoptic system.

They hypothesized that people would be able to perform well with both the 3D and the text system, but would feel less confident with the 3D system. The 3D system had most of the same "controls" (text query input, display of results), which they hypothesized would be used similarly to the text system. They hypothesized that people would feel less confident with their 3D results. Finally, they hypothesized there would be no significant differences in results across the different

search tasks, but anticipated slight ordering effects (increased performance over time with both systems).

University of Toronto

The primary objective of the Toronto group was to design a novel information exploration interface that combined multiple methods for accessing the content to accommodate the multiple perspectives that users bring to the task. Their work toward this goal was derived from the exploratory study done in conjunction with TREC 2001 Interactive Track. They had two goals for their TREC 2002 study:
1. Work toward a search workspace – a digital environment that contains a set of tools to aid users in exploring an information space
2. Develop a more cost-effective solution to information retrieval experimentation.

University of Glasgow

The Glasgow group assessed whether they could improve upon the limitations of the traditional elongated results list by an appropriate application of hierarchical clustering and summarization visualization techniques? Their current experimental system, provisionally named HuddleSearch, acted as an intermediate layer between the user and the provided Panoptic search engine. It used a newly developed clustering algorithm, which dynamically organizes the relevant documents into a traversable hierarchy of general to more specific clusters categories. They extended their TREC 2001 query-biased summarization tool to also allow the summarization of multiple documents, whereby a summary painted a caricature of the content of a cluster, rather than an individual document, thus allowing the user to provisionally judge a cluster's relevance prior to viewing its contents. The interaction between the user and the system was further developed by the aid of an information visualization tool.

From their initial analysis, they concluded that users do prefer the hierarchical clustering system to a list-based approach. Compared to the underlying Panoptic search engine, used as a baseline in their experimental design, the times taken to complete search tasks using their system clearly fell. In addition, the number of incomplete tasks was definitely reduced by the use of the experimental system.

Conclusions and Future Directions

While the track participants were pleased to be able to use Web data for experiments, a number of issues arose in this year's track. The main concern was the "open" .GOV collection, which made controlled experiments more challenging. Another problem was that the PDF files in the collection were proper consisted solely of text, which may be fine for batch experiments but is problematic for real users. Another general challenge for the track is the existence of high-quality commercial search engines such as Google, which make experimental systems frustrating for users when they do not perform as well as commercial products.

The track will undergo significant changes for TREC 2003. The track itself will become a subtrack of the Web track, and the current chair will be stepping down. Further details will be available on the TREC Web site.

References

1. Hersh WR, *Interactivity at the Text Retrieval Conference (TREC).* Information Processing and Management, 2001. 37: 365-366.
2. Wu M, Fuller M, and Wilkinson R. *Searcher performance in question answering. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2001. New Orleans, LA: ACM Press. 375-381.
3. Belkin NJ, et al., *Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval.* Information Processing and Management, 2000. 37: 403-434.
4. Hersh W, et al. *Do batch and user evaluations give the same results? Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2000. Athens, Greece: ACM Press. 17-24.
5. Hersh W and Over P, *SIGIR Workshop on Interactive Retrieval at TREC and Beyond*. 2000, SIGIR Forum, http://www.acm.org/sigir/forum/S2000/Interactive_report.pdf.