TREC-2001 Interactive Track Report

William Hersh
(hersh@ohsu.edu)
Division of Medical Informatics and 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

Motivating principles

In the TREC 2001 Interactive Track six research teams carried out observational studies which increased the realism of the searching by allowing the use of data and search systems/tools publicly accessible via the Internet. To the extent possible, searchers were allowed to choose tasks and systems/tools for accomplishing those tasks.

At the same time, the studies for TREC 2001 were designed to maximize the likelihood that groups would find in their observations the germ of a hypothesis they could test for TREC 2002. This suggested that there be restrictions - some across all sites, some only within a given site - to make it more likely that patterns would emerge. The restrictions were formalized in two sorts of guidelines: one set for all sites and another set that applied only within a site.

Cross-site guidelines

Each site observed as many searchers as possible and appropriate. A target number of 24 was suggested.

Each searcher worked in one or more of the following domains provided by the track to all sites:

- 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

Each searcher carried out four searches - two from a list of fully specified tasks provided to all sites and two for which only the format was predetermined but which were otherwise up to the site/searcher to create.

Each site collected a minimal standard set of data defined roughly by the track and covering searcher characteristics and satisfaction, effectiveness, and efficiency.

Each site collected at least the urls of all pages visited during all searches.

The only real submission required was the notebook paper, which was to include among other things a testable hypothesis for TREC-2002.

Tasks

Here are the eight fully specified tasks:

- Medical
  - Tell me three categories of people who should or should not get a flu shot and why.
  - Find a website likely to contain reliable information on the effect of second-hand smoke.
- Buying
  - Get two price quotes for a new digital camera (3 or more megapixels and 2x or more zoom).
  - Find two websites that allow people to buy soy milk online.
- Travel
  - I want to visit Antarctica. Find a website with information on organized tours/trips there.
  - Identify three interesting things to do during a weekend in Kyoto, Japan.
- Project
  - Find three articles that a high school student could use in writing a report on the Titannic
  - Tell me the name of a website where I can find material on global warming.

Here are the eight partially specified tasks:

- Medical
  - List two of the generally recommended treatments for _____.
  - Identify two pros or cons of taking large doses of _____.
- Buying
  - Name three features to consider in buying a(n) _____.
  - Find two websites that will let me buy a(n) _____ online.
- Travel
  - Identify three interesting places to visit in _____.
  - I'd like to go on a sailing vacation in _____, but I don't know how to sail. Tell me where can I get some information about organized sailing cruises in that area.
- Project
  - Find three different information sources that may be useful to a high school student in writing a biography of ________.
  - Locate a site with lots of information for a high school report on the history of _____.

Within-site guidelines

Within the cross-site guidelines, each site could impose further restrictions of its own choice on ALL its searchers to define an area of interest for observation - to be reported to the track before the observations begin. Each site could define its own time limits for searches. For example, a site could have imposed inclusive or exclusive restrictions on any (combinations) of the following: the choice/assignment of domain from the 4 provided, the data to be searched, the search system/tools to be used (e.g., search systems, meta-search systems, directories,...), functionality within a given search system/tool, the characteristics of searchers, the time allowed, the pre- search training provided, etc. Sites were also encouraged to coordinate their plans with other sites, form small teams sharing guidelines, etc. Each site evaluated their searches using any criteria defined in the cross-site guidelines plus any site specific evaluations. As part of the data analysis for TREC 2001, each site was to attempt to formulate a testable hypothesis for TREC 2002 and report this as part of the results for TREC 2001.

Overview of results

A total of six groups participated in this year's Interactive Track and submitted reports for the proceedings. Even though there was no official correct "answer" for any of the tasks, most groups attempted to assess some aspect of user searching performance, usually comparing two or more groups and/or systems. See each group's report for information about the formulation of testable hypotheses.

- Toms et al. [1] had 48 subjects who were given a choice of initiating the search with a query or with a selection of a category from a pre-defined list. Participants were also asked to phrase a selected number of their search queries in the form of a complete statement or question. The results showed that there was little effect of the task domain (medical, buying, travel, report) on the search outcome. There was a preference for the use of queries over categories when the semantics of the search task did not map well to one of the available catetories.

- Bhavhani [2] compared the searching behaviors of expert vs. non-expert searchers, with medical librarians and those experienced with on-line shopping performing both the flu-shot and camera tasks. There were substantial differences in how each group, with expertise in one area but not the other, performed the tasks. When searching in an area of expertise, the searchers tended to use more efficient, domain-specific resources and procedures, e.g., a site devoted to selling items of type X. When searching in an area outside their expertise they used more general-pupose methods (e.g. a general search engine to find a site for buying an X)

- Belkin et al. [3] looked at the role of increasing query length to see if it had any impact in task performance and/or interaction. Thirty-four subjects searched in one of two conditions: a "box" query input mode and a "line" query input mode. One-half of the subjects were instructed to enter their queries as complete sentences or questions; the other half as lists of words or phrases. The results showed that queries entered as questions or statements were longer than those entered as words or phrases (twice as long), that there was no difference in query length between the box and line modes, and that longer queries led to better performance.

- Hersh et al. [4] carried out a pure observational study, with users having their choice of which search engine or other resources to use. They measured time taken for searching, the number of pages viewed, satisfaction of users, and what topics users selected for their partially-formed searches. Their results showed that all the tasks took between six to ten minutes, with the buying task taking longest, followed by the medical, project, and travel tasks. User satisfaction was generally high, and the Google search engine was by far the most common starting point.

- Craswell et al. [5] assessed whether there was any correlation between delivery (searching/presentation) mechanisms and searching tasks. Their experiment involved three user interfaces and two types of searching tasks. The interfaces included a ranked list interface, a clustering interface, and an integrated interface with ranked list, clustering structure, and Expert Links. The two searching tasks were searching for an individual document and for a set of documents. Their results showed that subjects usually used only one interface regardless of the searching task. No delivery mechanism was found to be superior to any other for any particular task. The only difference noted was the time used to complete a search, which was less for the ranked list interface.

- White et al. [6] examined whether implicit feedback (where the system attempts to estimate what the user may be interested in) could act as a substitute for explicit feedback (where searchers explicitly mark documents relevant). They hypothesized that implicit and explicit feedback were interchangeable as sources of relevance information for relevance feedback, comparing the two approaches in terms of search effectiveness. No significant difference between the two approaches was found.

References

1. E.G. Toms, R.W. Kopak, J. Bartlett, L. Freund, Selecting Versus Describing: A Preliminary Analysis of the Efficacy of Categories in Exploring the Web, Proceedings of the Tenth Text REtrieval Conference (TREC 2001), in press.
2. S. K. Bhavhani, Important Cognitive Components of Domain-Specific Search Knowledge, Proceedings of the Tenth Text REtrieval Conference (TREC 2001), in press.
3. N.J. Belkin, C. Cool, J. Jeng, A. Keller, D. Kelly, J. Kim, H-J Lee, M-C Tang, X-J Yuan, Rutgers' TREC 2001 Interactive Track Experience, Proceedings of the Tenth Text REtrieval Conference (TREC 2001), in press.
4. W. Hersh, L. Sacherek, D. Olson, Observations of Searchers: OHSU TREC 2001 Interactive Track, Proceedings of the Tenth Text REtrieval Conference (TREC 2001), in press.
5. N. Craswell, D. Hawking, R. Wilkinson, M. Wu, TREC10 Web and Interactive Tracks at CSIRO, Proceedings of the Tenth Text REtrieval Conference (TREC 2001), in press.
6. R.W. White, J.M. Jose, I. Ruthven, Comparing Explicit and Implicit Feedback Techniques for Web Retrieval: TREC-10 Interactive Track Report, Proceedings of the Tenth Text REtrieval Conference (TREC 2001), in press.