Measuring Web Search Effectiveness: Rutgers at Interactive TREC

Nicholas J. Belkin, Gheorghe Muresan
{belkin, muresan}@scils.rutgers.edu
School of Library, Information and Library Studies
Rutgers University
What Should an IR System Do and What Should We Evaluate?

- Support the user’s
  - exploration of the problem domain,
  - clarification and refinement of the information need,
  - source selection, and
  - articulation of the information need in the appropriate input language or syntax;

- Use algorithms that retrieve as many relevant documents and as few non-relevant documents as possible;

- Support the user’s exploration of the retrieved documents, extraction of relevant information, and integration with the task at hand.
A View of Interactive IR

**INPUT**
- Problem definition
- Source selection
- Problem articulation

**OUTPUT**
- Examination of results
- Extraction of information
- Integration with overall task

Search Engine

Nicholas J. Belkin, Gheorghe Muresan
SCILS, Rutgers University
Some IR Evaluation Problems

- How best to evaluate performance of the system as a whole
- How to be realistic yet controlled
- How to gather sufficient and adequate data from which it is possible to generalize meaningfully
- How to tailor evaluation measures and methods to specific contexts and tasks
Our Foci Today

- Measures and methods appropriate for investigating effectiveness of interactive IR on the Web

- Formative (analytic) evaluation

- Based on our experiences in the TREC Interactive Track and TREC Web Interactive Sub-track
## A Classification of IR Evaluation Measures

<table>
<thead>
<tr>
<th>Task Specificity</th>
<th>General</th>
<th>Task-specific</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interactivity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-interactive (laboratory evaluation of the retrieval algorithm)</td>
<td>Recall, precision, E, F; Expected search length</td>
<td>Question answering: mean reciprocal rank of the correct answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Filtering: utility</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic distillation: coverage and accuracy</td>
</tr>
<tr>
<td>Interactive (evaluation of the interaction process and outcome)</td>
<td><em>User satisfaction, learnability, ease of use</em></td>
<td>Perception of task performance</td>
</tr>
<tr>
<td></td>
<td>User effort (clicks, iterations, scrolling, documents seen, viewed or read)</td>
<td>Aspect retrieval: aspectual recall, number of saved documents</td>
</tr>
<tr>
<td></td>
<td>Efficiency: time to complete task, precision at N seen</td>
<td>Question answering: completeness and correctness of answer</td>
</tr>
<tr>
<td></td>
<td>Expected search length</td>
<td>Topic distillation: coverage and accuracy</td>
</tr>
</tbody>
</table>
Lessons Learned from the TREC Experience

- IR, especially Web IR, is inherently interactive,
  - measures of search effectiveness alone are insufficient

- Information seeking is engaged in for many different purposes, in many different contexts, to accomplish many different tasks
  - one (or one set of) measure(s) for evaluating IR in general is a Chimera
Lessons Learned from the TREC Experience

- It may not be a good idea to rely on external “objective” judgments for evaluation purposes.

- Experimental methods can be used successfully in user-centered evaluation of interactive IR.
Some Conclusions or Recommendations

- *Perceptions* of performance are as important as “objective” measures; both should be interpreted w.r.t. measures of the search process.

- Different measures need to be established w.r.t. goals of different tasks.

- Specific experimental tasks should be designed so that the subjects’ performance in the task, and the subjects’ own evaluation of performance, are the criteria for the evaluation measures.