Method Bias? The Effects of Performance Feedback on Users’ Evaluations of an Interactive IR System

or

What we did last semester in the IIR Seminar


CRADLE | 04 April 2008
Motivation

• Let’s start with the traditional IIR evaluation model (a lá TREC)
• What’s wrong with this model?
  – Relevance Assessments
  – Assigned Topics
  – Not much variability in Tasks
  – …
Motivation

• How can users be expected to evaluate a system when they have no idea how well they performed?
  – Some problems with ‘perceptions’
  – Users like everything
Research Question

• How does providing feedback to users about their performances for high-recall tasks affect their evaluations of an experimental IIR system?

• “New” Evaluation Method
• Feedback: Recall (relevant documents found/relevant documents in corpus)
Feedback Conditions

- **No Feedback (NF):** users are not provided with any feedback about their performances (baseline)
- **Actual Feedback (AF):** users are told their real performances
- **Good Feedback (GF):** users are deceived and told they performed very well (92%)
- **Poor Feedback (PF):** users are deceived and told they performed very poorly (12%)
Expectations

Poor Feedback < Actual Feedback < No Feedback < Good Feedback
Method

• XRF experimental IR system + TREC collection + Lemur (BM25)
• 60 undergraduate subjects assigned randomly to condition (between subjects)
• Three recall-oriented search topics followed by Post-Task Questionnaires
• For three conditions (PF, AF, and GF) performance feedback is provided before completing Exit Questionnaire
Example Topic

[simulated work task] + Your goal is to identify as many different developments in robotic technology and their uses as possible, and to find as much information about each development and its uses as possible.
“Flavor” of Questionnaires

- Post-Task Questionnaire (6 items)
  - Familiarity, Easy to search, Satisfaction with search results, Confidence, Enough time, Satisfaction with performance
  - 7-pt scale: 1=not at all, 4=somewhat, 7=extremely
“Flavor” of Questionnaires

- Exit Questionnaire (13 items)
  - Easy to learn, Inconsistencies, Easy to query, Easy to navigate, Similarity of search methods, Color-coding, Easy to use, Accomplish, Easy to find relevant documents, Easy to understand why documents were retrieved, Various functions well-integrated, Overall effectiveness, Overall satisfaction
  - 7-pt scale: 1=strongly disagree, 7=strongly agree
Results

• Post-Task Questionnaire Responses
  – For most items (except familiarity), ratings were higher than the scale mid-point
  – For all items, there were no significant differences in responses according to condition
  – Overall, things seems to be going okay ...

<table>
<thead>
<tr>
<th>Feedback Condition</th>
<th>Poor</th>
<th>Actual</th>
<th>No</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfaction</td>
<td>4.69 (1.20)</td>
<td>5.22 (1.19)</td>
<td>4.89 (1.01)</td>
<td>4.82 (1.28)</td>
</tr>
</tbody>
</table>
Exit Questionnaire #13: Satisfaction

$F(3,59)=9.54, p<.01$

PF < NF, GF
AF < GF

Effect Size = .34
Pre- & Post-Feedback Satisfaction

Feedback Condition

NF: $t(14) = 1.89^*$
AF: $t(14) = -3.14^{**}$
GF: $t(14) = 6.29^{**}$
PF: $t(14) = 3.58^{**}$

*ns
**$p<.01$
Exit Questionnaire Responses

Average Score of subjects in the No Feedback Condition for each Question is set to 0.
A Closer Look at Three Other Questions

2. I didn’t notice any inconsistencies when I used the system.

3. It was easy to pose queries to the system.

4. It was easy to navigate the search results.
Exit Questionnaire #2

$F(3,59)=3.60, \ p=.02$

AF < NF

Effect Size = .16
Exit Questionnaire #3 and #4

\[ F(3,59)=4.50, \ p=.01 \]

PF < NF, GF

Effect Size = .19

\[ F(3,59)=2.99, \ p=.04 \]

ns

Effect Size = .14
What do we make of these results?

• We consider several explanations:
  – Inaccurate perceptions of performance
  – Inflation bias
  – Attribution bias
Inaccurate Perceptions

• Subjects’ perceptions of how well they performed were inaccurate and providing them with information about this helped them to make more accurate evaluations

• No strong relationship between objective and subject measures

• Subjects in the No Feedback condition gave the system some of the highest ratings despite being some of the worst performers

• Pre- and post-test changes in ‘inconsistencies’ question
Inflation Bias

- Subjects’ initial ratings were inflated and providing them with information about their performances motivated them to make more critical evaluations.
- Several items affected by feedback condition were about interface features.
- Regardless of feedback condition, average ratings were still above the scale mid-point.
Attribution Bias

- Subjects in the Poor and Actual Feedback conditions blamed the computer for their ‘failures’
- Attribute theory is often used to explain how people perceive the causes of success and failure and how people attribute blame
- Does IIR create an equal-opportunity ‘blame’ scenario? Who is responsible for the outcome?
- Our study was not set-up to really support exploration of this explanation
Some Limitations

- Subjects’ interpretations of performance measures (i.e., 30% = F)
- Validity and reliability of questionnaire items
- Expectations about the relationship between objective and subjective evaluation measures
Conclusions

• Lots of undergraduates are no-shows
• Subjects’ evaluation behaviors are brittle and susceptible to slight changes in evaluation method
• Researchers should provide users with feedback about their performances when this information is available (?)
• Researchers should at least start to question and evaluate method variance
Teaching + Research

• Start with something you *already* have available and a *simple* research question.
• Ideally you have published a paper about this topic. Ask students to read this and discuss it in class.
• Spend class time discussing and refining instruments and method
• Provide lots of training & structure
• *(Where’d I get the $600?)*
Teaching + Research

- As data are being collected, discuss experiences (research therapy)
- Provide a template for writing report describing data *each student* collected
- Discuss individual findings and present overall key findings
- If possible, integrate some of material from the reports into the final manuscript
- Master’s papers
Fifteen catastrophic hurricanes, floods and storms cost worldwide insurers more than \$100 billion (Pounds 50 billion) since a period of weather extremes set in 10 years ago, according to an article in the latest World Watch Institute’s journal.

In 1992, Hurricane Andrew struck Florida and set a new record for damages at \$20 billion. The Mississippi floods in 1993 cost \$13 billion. Europe was hit by four severe winter storms in 1990 which accumulated damages of \$10 billion. Japan was struck in 1991 by Typhoon Vincenzo with nearly \$3 billion in damages.

As the damages mount, insurers have begun to take seriously the growing threat of global warming observed by many scientists. The main reason is that the warming, spurred by ‘greenhouse gases’, produced by burning fossil fuels, could seriously disrupt the world’s atmospheric and oceanic systems.

Lack of agreement in the scientific community has made the insurers wary. But their interest is being applauded by environmentalists who see the insurers as a ‘potential public watchdog to the power of the oil and coal industries in the global warming debate.’

Christopher Flinn, author of the World Watch article, is urging the insurers to enter the struggle over climate policy. ‘The industry is capable of doing battle with the likes of the fossil fuel lobby. But the insurance industry, he says, is a worldwide body the two are of roughly comparable size and potential political clout.

The insurance industry could, for example, push governments to tighten energy-efficiency standards for new buildings. It could also lobby for a stronger global climate pact.

It could also use its investment capacity. ‘If companies were to dump some of their stocks in oil and coal companies or advocated (made) some of their funds in new, less carbon-intensive energy technologies forming a sort of climate-watcher fund, insurance companies could spur the development of a less threatening energy system,’ says Flinn.

Unless the industry begins to use its clout in the struggle over climate policy, its future is likely to be stormy indeed, said Flinn.
Basic Protocol

START

Greeting & Consent
System Tutorial

[Repeat for N Tasks]

Introduce Task
Subject Searches
Post-Task Questionnaire

• Observation
• System Logs

Exit Questionnaire
Exit Interview

END
START

Greeting & Consent
System Tutorial

[Repeat for N Tasks]

Introduce Task
Subject Searches
Post-Task Questionnaire

Performance Feedback

END

• Observation
• System Logs

Exit Questionnaire
TREC
[Text REtrieval Conference]

It’s not this ...
What is TREC?

- TREC is a workshop series sponsored by the National Institute of Standards and Technology (NIST) and the US Department of Defense.
- It's purpose is to build infrastructure for large-scale evaluation of text retrieval technology.
- TREC collections and evaluation measures are the *de facto* standard for evaluation in IR.
- TREC is comprised of different tracks each of which focuses on different issues (e.g., question answering, filtering).
Table 1.1
Number of participants per track and total number of distinct participants in each TREC

<table>
<thead>
<tr>
<th>Track</th>
<th>TREC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>92</td>
</tr>
<tr>
<td>Ad hoc</td>
<td>18</td>
</tr>
<tr>
<td>Routing</td>
<td>16</td>
</tr>
<tr>
<td>Interactive</td>
<td>—</td>
</tr>
<tr>
<td>Spanish</td>
<td>—</td>
</tr>
<tr>
<td>Confusion</td>
<td>—</td>
</tr>
<tr>
<td>Database merging</td>
<td>—</td>
</tr>
<tr>
<td>filtering</td>
<td>—</td>
</tr>
<tr>
<td>Chinese</td>
<td>—</td>
</tr>
<tr>
<td>NLP</td>
<td>—</td>
</tr>
<tr>
<td>Speech</td>
<td>—</td>
</tr>
<tr>
<td>Cross-language</td>
<td>—</td>
</tr>
<tr>
<td>High precision</td>
<td>—</td>
</tr>
<tr>
<td>Very large corpus</td>
<td>—</td>
</tr>
<tr>
<td>Query</td>
<td>—</td>
</tr>
<tr>
<td>Question answering</td>
<td>—</td>
</tr>
<tr>
<td>Web</td>
<td>—</td>
</tr>
<tr>
<td>Video</td>
<td>—</td>
</tr>
<tr>
<td>Novelty</td>
<td>—</td>
</tr>
<tr>
<td>Genome</td>
<td>—</td>
</tr>
<tr>
<td>HARD</td>
<td>—</td>
</tr>
<tr>
<td>Robust</td>
<td>—</td>
</tr>
<tr>
<td>Total participants</td>
<td>25</td>
</tr>
</tbody>
</table>
Central to each TREC Track is a collection, which consists of three major components:

1. A corpus of documents (typically newswire)
2. A set of information needs (called *topics*)
3. A set of relevance judgments.

Each Track also adopts particular evaluation measures:
- Precision and Recall; F-measure
- Average Precision (AP) and Mean AP (MAP)
Learn more about TREC