Cognitive Topics in User Modelling in Interactive Information Retrieval

Alan Medlar (alan.j.medlar@helsinki.fi)
Dorota Glowacka (glowacka@cs.helsinki.fi)
Organisation and assessment

• This seminar is worth 5 credits

• Duration: teaching periods 1 and 2

• Aim: introduce interactive information retrieval and cognitive modelling

• Assessment:
  • 8-10 page report (+ references) on using cognitive modelling in IIR
  • a presentation on the topic of the report

• Topics in IIR and cognitive modelling can be chosen freely, but we will provide suggestions
Schedule

- Changes will appear on the course webpage
- 04.09.19 Lecture 1: Introduction to IR and IIR
- 11.09.19 Lecture 2: Cognitive modelling
- 25.09.19 Deadline for topic selection (title + 3 papers min.)
- 09.10.19 Presentation of chosen topic (5 mins, 5 slides)
- 30.10.19 Feedback session
- 20.11.19 Final presentations (20 mins, 20 slides) - if necessary
- 27.11.19 Final presentations (20 mins, 20 slides)
- 11.12.19 Deadline for final paper submission
Essay structure

- Essay will have 3 sections:
  
  - **An IIR component (or search task)** (e.g. ranking, relevance feedback, implicit relevance feedback) - what does it do? how is it implemented? how is its effectiveness validated?
  
  - **A cognitive process** (e.g. categorisation, decision making, implicit learning) - what does it study? describe the model, what type of experiment is used to gather data?
  
  - **Cognitive modelling in IIR** (e.g. modelling relevance feedback as a categorisation process) - sketch an experimental design, what old results can be replicated? what new results will we get?
Information Retrieval (IR)

- Definition (Introduction to Information Retrieval, Manning, Raghavan and Schütze, 2010):

"Information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."
IR example
Laboratory model of IR
IR Evaluation (1)

- Based on the Cranfield experiments (1966)

- Assumes that the **relevance** of retrieved documents is a good proxy for whether the IR system satisfies users' **information needs**

- Requirements:
  - Document corpus
  - Information needs (queries)
  - Relevance judgments (binary assessment of relevant/not relevant for query-document pairs)
IR Evaluation (2)

• Precision and recall used to evaluate unranked search results.

• Tradeoff between precision and recall - as the number of search results increases, precision decreases, but recall increases (on average).
IR Evaluation (3)

• Precision and recall does not reflect the efficacy of IR systems when search results are **ranked**

• Evaluation metrics for ranked search results:
  
  • Precision-recall curve
  
  • Precision@K (P@K)
  
  • Mean average precision (MAP)
  
  • Discounted cumulative gain (DCG)

\[
\text{DCG}_p = \sum_{i=1}^{p} \frac{2^{\text{rel}_i} - 1}{\log_2(i + 1)}
\]
Models of search

• Classic IR
  • Content-related search in unstructured documents
  • System-oriented view
  • Static information needs
Classical search process model

1. Information Need
2. Query
3. Send to System
4. Receive Results
5. Evaluate Results
6. Reformulate
   - If Yes: Go to 7
   - If No: Go to 3
7. Done?
   - If Yes: Stop
   - If No: Go to 2
Empirical studies

- Search consists of a sequence of connected, but different searches
- Search results trigger new searches, the task context remains constant
- Goal of search is to accumulate information and learn about the search topic
Models of search

- Classic IR
  - Content-related search in unstructured documents
  - System-oriented view
  - **Static** information needs
- Interactive IR
  - Focus on user interaction with information system
  - **Dynamic** information needs
Berry-picking model of information seeking

Marchionini's search activities

Interactive Information Retrieval (IIR)

- Definition (Methods for evaluating interactive information retrieval systems with users, Kelly, 2009):

"In interactive information retrieval (IIR), users are typically studied along with their interactions with systems and information. While classic IR studies abstract humans out of the evaluation model, **IIR focuses on users’ behaviors and experiences**—including physical, cognitive and affective—and the interactions that occur between users and systems, and users and information."
Relevance in IIR

- **CANNOT** assume that the *relevance* of retrieved documents is a good proxy for whether the IIR system satisfies users' *information needs*

- Relevance assumed to be *subjective*: related to the user's knowledge, interests, etc.

- Cannot use relevance judgments, no substitute for real users!
IIR example (mid90s)

IIR example (2016)

- PULP video
IIR Evaluation

• "... there is no strong evaluation or experimental framework for IIR evaluations as there is for IR studies."

• Study design (search task, within-subjects vs. between-subjects)

• Measurements (think-aloud, observation, logging, questionnaires, semi-structured interviews)

• Data analysis (statistical tests, non-parametric tests, repeated-measures ANOVA, regression)
Search behavior measures

- Search behavior measures are logged by the interface or the backend
- Example from Kelly et al. 2015

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Queries</td>
<td>Total number of unique queries submitted by a participant when completing a task.</td>
</tr>
<tr>
<td>Query length</td>
<td>Average number of query terms in all unique queries issued for a task.</td>
</tr>
<tr>
<td>Unique query terms</td>
<td>Total number of unique query terms used by a participant when completing a task.</td>
</tr>
<tr>
<td>SERP clicks</td>
<td>Total number of clicks participants made on SERPs.</td>
</tr>
<tr>
<td>URLs visited</td>
<td>Total number of unique URLs visited by participants (includes URLs accessed directly and indirectly via SERP).</td>
</tr>
<tr>
<td>Queries w/o SERP clicks</td>
<td>Total number of unique queries where participants did not click on the search engine results page (SERP).</td>
</tr>
<tr>
<td>Time to completion</td>
<td>The amount of time (in seconds) participants spent completing search tasks.</td>
</tr>
<tr>
<td>SERP dwell time</td>
<td>Average time spent between issuing a new query and clicking on the first search result (in seconds).</td>
</tr>
<tr>
<td>Query diversity</td>
<td>Number of queries issued that were not issued by another participant completing the exact same task.</td>
</tr>
<tr>
<td>Query term diversity</td>
<td>Number of query terms used that were not used by another participant completing the exact same task.</td>
</tr>
<tr>
<td>URL diversity</td>
<td>Number of URLs visited that were not visited by another participant completing the exact same task.</td>
</tr>
</tbody>
</table>

### Questionnaires

**Table 3. Pre-Task Questionnaire Items**

<table>
<thead>
<tr>
<th>Interest &amp; Knowledge</th>
<th>How interested are you to learn more about the topic of this task?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>How many times have you searched for information about this task?</td>
</tr>
<tr>
<td></td>
<td>How much do you know about the topic of the task?</td>
</tr>
<tr>
<td>Task Complexity</td>
<td>How defined is this task in terms of the types of information needed to complete it?</td>
</tr>
<tr>
<td></td>
<td>How defined is this task in terms of the steps required to complete it?</td>
</tr>
<tr>
<td></td>
<td>How defined is this task in terms of its expected solution?</td>
</tr>
<tr>
<td>Expected Task Difficulty</td>
<td>How difficult do you think it will be to search for information for this task using a search engine?</td>
</tr>
<tr>
<td></td>
<td>How difficult do you think it will be to understand the information the search engine finds?</td>
</tr>
<tr>
<td></td>
<td>How difficult do you think it will be to decide if the information the search engine finds is useful for completing the task?</td>
</tr>
<tr>
<td></td>
<td>How difficult do you think it will be to integrate the information the search engine finds?</td>
</tr>
<tr>
<td></td>
<td>How difficult do you think it will be to determine when you have enough information to finish the task?</td>
</tr>
</tbody>
</table>

**Table 4. Post-Task Questionnaire Items**

<table>
<thead>
<tr>
<th>Engagement</th>
<th>How enjoyable was it to do this task?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>How engaging did you find this task?</td>
</tr>
<tr>
<td></td>
<td>How difficult was it to concentrate while you were doing this task?</td>
</tr>
<tr>
<td>Interest</td>
<td>How much did your interest in the task increase as you searched?</td>
</tr>
<tr>
<td></td>
<td>How much did your knowledge of the task increase as you searched?</td>
</tr>
<tr>
<td>Experienced Task Difficulty</td>
<td>Same five items from Table 3 except items started with,</td>
</tr>
<tr>
<td></td>
<td>“How difficult was it to ...”</td>
</tr>
<tr>
<td>Overall Difficulty</td>
<td>Overall, how difficult was this task?</td>
</tr>
<tr>
<td>Overall Satisfaction</td>
<td>Overall, how satisfied are you with your solution to this task?</td>
</tr>
<tr>
<td></td>
<td>Overall, how satisfied are you with the search strategy you took to solve this task?</td>
</tr>
</tbody>
</table>

Research Questions, Hypotheses, and Theory. Explicit research questions were found in 19.3% of the studies ($n = 29$), explicit hypothesis were found in 10.7% ($n = 16$) of the studies, and both a research question and a hypothesis were found in 4.7% of the studies ($n = 7$). In $65.3\%$ ($n = 98$) of the studies, there was neither an explicitly stated research question nor hypothesis.

In the majority of studies ($n = 57, 45\%$), ANOVA was used as the method of analysis. This was followed by $t$ test ($n = 33, 26\%$), Mann-Whitey ($n = 11, 9\%$), chi-square ($n = 8, 6\%$), and Wilcoxon signed-rank test ($n = 6, 5\%$). Correlation, Kruskal-Wallis and factor analysis were observed in fewer than $5\%$ of the articles. Fifteen percent ($n = 19$) of the articles presented only descriptive statistics, while $9\%$ ($n = 11$) did not provide any indication of which type of analysis was used, despite claiming statistically significant results or presenting probability values. Almost all the analyses were performed variable-by-variable and were conducted to compare the systems. Only a small percentage of articles described statistical analyses that attempted to model performance using multiple input variables ($n = 6, 5\%$).
Researcher degrees-of-freedom

Coined the term "researcher degrees-of-freedom"

Choice between two dependent variables nearly doubles false positive rate
Reading


Next lecture: Cognitive models and applications