Obtrusiveness and Relevance Assessment in Interactive XML IR Experiments

Birger Larsen, Tassos Tombros and Saadia Malik

Royal School of Library and Information Science
Queen Mary University of London
UNIVERSITÄT DUISBURG ESSEN

Outline

- The Interactive Track at INEX
- Relevance assessments
- 2004 compromise
- Alternatives
- Conclusions and discussion points

The INEX Interactive Track

- First time at INEX2004, 10 active groups
- Purposes
  - Investigation of user interaction with XML systems
  - Development of XML IR approaches that are effective in user-based environments
  - Feed information back to the ad hoc track
- Few groups had systems → collaborative effort
  - Baseline system used by all, 8 searchers per site
  - Used CO topics only (NEXI too complex for users)

Relevance Assessments

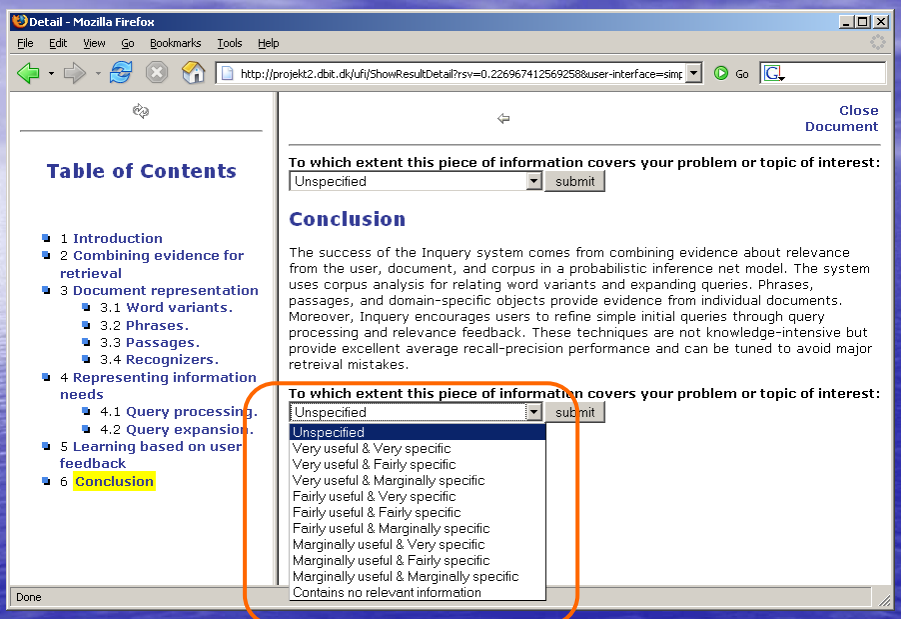
- What relevance assessments to use?
- Research questions to answer
  - Granularity issues: Skimming of larger elements vs. direct presentation of smaller ones
  - Do users gain anything by browsing the document structure?
  - Sensitivity to redundant information and overlaps
  - ... is element retrieval ultimately of value to users

Relevance Assessments

- Ideally, we'd like the test persons to assess at least the following for each viewed element
  - The amount of relevant vs. irrelevant information (~ *Specificity*)
  - How much of the work task that can be solved by the element (~ *Exhaustiveness*)
  - *Redundancy* in results
  - Overall *usefulness/pertinence*
- Problem
  - High cognitive load for test persons
  - "Natural" browsing behaviour will be effected
  - May even be experienced as *obtrusive*

2004 compromise

- Attempt to maintain a certain level of comparability with ad hoc track
  - 2 relevance dimensions, with graded assessments merged into a single 10-point combined scale
- Drawbacks
  - Not all viewed elements were assessed (~60 %)
  - Fairly obtrusive...
  - Difficulties in understanding the scale?
  - Not easy to infer reasons behind assessments



Alternatives

1. No interactive assessments, use ad hoc assessments
   - Pros: Easy access to existing assessments, minimum strain on test persons
   - Cons: Fundamentally opposed to the idea of interactive studies
2. Use implicit indicators (time spent, scrolling, eye movements)
   - Pros: Minimum strain
   - Cons: Hard to relate relevance to implicit indicators, especially to specific levels of exhaustiveness and specificity; may be difficult to gather and analyse data

Alternatives

3. Less comprehensive, but explicit assessments (e.g., bookmarking)
   - Pros: Not very obtrusive, part of natural search behaviour?
   - Cons: Almost no indication of why an element was bookmarked; many un-assessed elements
4. Comprehensive assessments with simple relevance scale
   - Pros: Assessments can be completed faster with a simple scale
   - Cons: No indication of why an element is relevant; fairly obtrusive

Alternatives

5. Talk-aloud protocols obtained during interaction
   - Pros: Information on why elements are relevant
   - Cons: Obtrusive?; labour intensive data recording & analysis
6. 'Talk-after' interviews, using recorded video/ eye movements + bookmarking/simple scale
   - Pros: Not obtrusive (?); information on why elements are relevant
   - Cons: labour intensive (conducting experiments, recording & analysing data); need for special equipment

Conclusions

- We'd like a wide range of aspects assessed for each viewed element
- …but this may prevent natural behaviour, be obtrusive, and undermine the purpose of interactive studies (test persons = a second team of assessors?)

Points for discussion

- Is bookmarking/simple scale + talk-after interviews a better alternative for getting at the *why* of element retrieval?
- Is this setting feasible in a distributed experiment and worth the extra cost?
- Any other data collection methods?