Panel

Issues in the evaluation of AIR systems?

Panelists: Kalervo Järvelin, Mark Sanderson, David Harper, Nick Belkin

Kal

- Standard approach to test collections, components: collection, task, assessments
- We could have different kinds of assessments, topics, more varied collections
- What are test collections used for?
  - Performance of the engine, focus on improving recall/precision power of engine
  - Interaction power?
  - Task/outcome evaluation
- Shared datasets vs. Standard test collections
  - Interaction database might not help us determine performance
- Users are irrational
  - What are real user interaction tests?
  - How irrational are real users?
- Keith: there are models of irrational behaviour

Mark

- Ellen is right that test collections are very good (best possibility)
- However, multiple relevance judgements has a potential to affect ranking algorithms
- There are solutions to some aspects to test AIR
  - User logs
  - Modelling/prediciting user behaviour
  - No need for terrabytes of logs
  - Logs seem to be usable for other contexts as well
- Research 2.0 (for usability testing)
  - Deploy system on large scale on network (large user base), analyse logs

David

- Good scientist should question everything in order to advance field
  - Single user relevance assessments
- Not enough to represent one user
- Timestamps, user queries, set of documents and actions to represent user interaction
- Proposes new study/challenge: Fix the engine and make a study where people only adapt the interface (or tasks, ...)
- What are interesting measures?
  - Cognitive load
  - Recall velocity (how fast is recall changing)

Nick

- What do we mean by a test collection for AIR?
  - Is TREC the only possibility?
  - Should it be thought of as a shared resource for investigating AIR?
- TREC-like test collection might not be suitable to study AIR factors
- What could a shared resource look like?
  - What are the facets of adaptations? What are the factors we feel are significant in helping a system to adapt to a user (context, situation)?
  - What kinds of information should be included?
- We don‘t have a strong handle on what facets are significant yet? -> aim to be as inclusive as possbile
- What are the actual/essential tasks that adaptation is supposed to be addressing?
  - What can be measures/understood via a test collection/shared resource?
- Start of building shared resource is to try to collect the data collected by participants of TREC interactive track (TREC 3-9)
  - User characteristics (eg knowledge of topic)
  - User satisfaction
  - User actions
  - Very little of these data has ever been analysed
- We don‘t need terrabytes of logs of different people, but terrabytes of data *about* people
- We should share logging tools, define logging standards

Discussion

- Leif: How can user logs help to create user models for simulated user studies?
  - Kal: Importance is to model the most important factors
  - David: New datasets would provide additional information to get better descriptions of the factors, help to improve realism of evaluation.
- Leif: Is simulated testing like animal testing in the process (referring to Noriko‘s talk)?
  - When work task is considered, real people have to be considered
- Keith: Should we try and encourage a group of people to take on the project of specifying what a test collection ought to look like? (design and specification, recommendations rather than collecting such data)
- Keith: Ellen mentioned that when interaction increases, variability increases. So we need to define what the data should look like, what will be its limitations
- Kal: We have to be very clear what the research question is
- Nick: National Science Foundation (NSF) might have funding options for supporting such a project, possibly even a joint program with EU
- Kal: Measure the quality of results, task performance time; need to define appropraite measurements, measurements might have tobe adapted for different user populations
- Diane: Importance of maximising diversity in user sampling (user characteristics, tasks) when creating AIR test collection, need to know what the limitations of the collection are
- Kal: Risk in AIR test collection might be that we find out that IR engine doesn‘t matter
- Kal: Risk is that we only look at average performance measure, rather than try to understand what the system does
- Nick: Measure of informativeness, also should try to relate use of documents
- Diane: Affective components are important, JASIST publicaton

Wrap up (Keith)

- Lot of work needs to be done to define Adaptive Information Retrieval and suitable evaluation methodologies