Evaluating Information Access Systems

Purpose of the visit;

People engage in information seeking (in our case, enter into interaction with an IR system) in order to achieve a goal or accomplish a task. At the most general level, the goal of the IR system is to support the person(s) engaged in information seeking in achieving their goal, accomplishing their task. Although some kinds of goals or tasks may be achieved through a single query-response cycle in an IR system, there is substantial evidence that there are many instances, in Web searching and in other types of IR systems, of search sessions which comprise of two or more such cycles, with a variety of behaviors exhibited by the information seekers in the intervals between query entry.

Current methods and measures for evaluation of IR system performance are limited to evaluation of a single-query search session, and are not at all suited to the evaluation of system performance with respect to search sessions which have more than one query. This holds not only with respect to performance of the system for the whole search session, but for evaluation of the support the system offers for the separate components or stages throughout the search session. Multiple query sessions necessitate that all components of search interaction, including the searcher's underlying motivating task, the search task, the searcher him/herself, the interface, and the search engine, are taken into account. Currently too little is known about their contribution to search success and about their interaction. For these reasons, it is necessary to develop evaluation criteria, measures and methods which are appropriate for evaluating IR system support for entire search sessions, and their elements. In particular, it is necessary to develop criteria, measures and methods which are predicated upon, and sensitive to, achievement of the different types of goals in tasks of different types which lead people to engage in information seeking behavior.

Whole-session evaluation is complicated due to the many factors involved. Therefore several quite different approaches and frameworks are required to solve the problems. These include observational studies, controlled user studies, simulation studies, and living labs.

The purpose of the visit was to work together with leading researchers in the field of information retrieval evaluation on (a) identifying the most significant problems in accomplishing whole-session evaluation of access systems, (b) propose some means of addressing these problems, (c) suggest some possible evaluation frameworks for accomplishing whole-session evaluation. The set of the declared end goals of the meeting was:

- A document outlining:
  - Need for whole-session evaluation;
  - Problems facing whole-session evaluation;
  - Potential approaches to those problems, with pluses, minuses and contexts of application;
  - Proposals of research directions in implementing whole-session evaluation

- A set of possible frameworks in which whole-session evaluation could take place
- A proposal to TREC (or some other evaluation forum) for a Whole-Session Evaluation Track in 2013
- Establishment of a research community to share data and experiences

Description of the work carried out during the visit;

The the meeting was constructed around a number of “food-for-thought” talks and break out sessions.

There were three such “food-for-thought” talks, the first was on “User task understanding: a web search engine perspective” presented by Peter Bailey from Microsoft Bing; the second on “TREC Session Track” for which I have been a coordinator for the past 3 years and it was presented by myself describing the vision of the track, the goal of the track, the current evaluation framework with the pros and the cons and discussing potential extensions with the researchers participating in the meeting; the last talk was on “Models of Search Behavior” presented by Diane Kelly from the University of North Carolina, USA.

Based on the “food-for-thought” talks and a set of papers the participants of the meeting identified there was a follow-up breakout session around three themes: (a) ideal session evaluation, (b) quantitative modeling of sessions, and (c) identifying tasks & goals or particular domains. I joined the discussion in the ideal session evaluation breakout group. Members of the different groups summarized and presented the discussions held in each of the groups. Based on these presentations a new set of breakout groups was formed around three themes again: (a) simulation, (b) shared experimental materials and methods, and (c) detecting session types/components/boundaries. I joined the discussion in the simulation breakout group.

Description of the main results obtained;

A large number of aspects of whole-session evaluation was examined and thoroughly discussed. A document discussing the main results of the meeting was drafted and will be submitted for publication in the SIGIR Forum. In what follows I describe some of the key points discussed in the two breakout groups I participated and which focused on (a) what consists an ideal test collection for the purpose of whole-session evaluation, and (b) how can we use simulations as a framework for whole-session evaluation.

The focus of the first breakout group was what consists **an ideal whole-session evaluation test collection**; that is how can we assemble a collection of document; how can we build queries and topics that could lead to more than a single query sessions; and what would be a set of evaluation measures that would better capture success over sessions?

Some of the question raised and thoroughly discussed around the concept of ideal whole-session evaluation test collections are the following:

As non-independence of actions within a whole-session task is one of the key open problems for whole-session evaluation, the first important research direction is to identify under what conditions *independence assumptions are violated*. There may indeed be a number of tasks or scenarios in which independence assumptions hold, and therefore the best system is the one that returns the session-context-free best results at every point in time. Can we separate these scenarios from others in which handling dependencies is crucial?

Another important research direction involves simulation and, in particular, *user simulation models*. Is it possible to validate these simulations by linking them with user studies? Is it also possible to link small scale user studies with large scale logs and infer user type or nearest user model to an observable set of behaviors?

Furthermore, there are not only a number of different user types, but a number of *different task types* as well. These include exploration, planning (with transitions and interleaving between subtasks and subtopics), classic finding and re-finding, learning, and so on. Depending on task type, different sorts of system support are required to support whole-session interactivity. One future research direction is to discover whether certain types of interactivity, whether supported by adaptive interfaces or by the nature of the system-provided results themselves, provide more or less utility when it comes to supporting various tasks.

Finally, there are a number of research directions in the standard, non-interactive domain that require integration and rethinking within a whole-session perspective. For example, information retrieval classically defines the basic unit of retrieval as the document. It might be more informative to redefine this in terms of information units, as these correlate better with concepts of novelty and redundancy. Eyetracking is another important source of evidence, as they allow us to drop assumptions such as entire snippets/documents being read by the user; and linear traversal (user scans a ranked list from top to bottom). How the relationship between eyetracking and evaluation change when integrated into a whole-session framework?

The focus of the second breakout group was whether **simulating user interactions** with a retrieval system could offer a framework that allows for whole-session evaluation. Throughout this study session a number of potential simulation approaches were identified and discussed.

Simulations with Limited Interaction (Batch mode):

One possible approach, that takes a step towards evaluating the whole session, is to extend and augment the traditional batch TREC evaluations. To simulate interaction of the first few queries within a session a batch approach can be taken by following a process like the following. Participants are provided with an initial query i.e. the query to start the session, along with a set of possible subsequent queries. In response, participants will need to produce rankings based on the initial query, and the initial query plus the subsequent query.

The assumption is that the subsequent query was issued after the initial query – and thus part of a session. However, the system receives no feedback regarding what documents in the initial ranking were clicked, etc (which is a limitation of the approach, but makes it possible to run in a batch mode). Participants can also provide additional subsequent queries, and results, which are added a "query pool". To evaluate the systems, simulated users are defined, and depending on the initial set of results, the simulated user will then select a subsequent query from the query pool of subsequent queries. Different simulated users will of course select different next queries. This is where the interaction of the initial part of the session is simulated - essentially boiled down to query refinement. While this approach follows on from standard TREC model, it only provides a very limited interaction, with a fixed abstraction over the interface, etc, it does provide a tractable approach that could be implemented as an evaluation track.

Simulations with Interaction on the fly (interactive mode):
A more ambitious approach would be to perform a fully interactive simulation of a session (or part thereof) where simulated users interact with an IR system. That is a simulated user would, given an information need and background, formulate a query and submit the query to the system. The system would respond with an abstracted interface / search engine result page which could include result snippets, facets, query suggestions, etc. The simulated user would then, given this response, perform some interactions with the page and formulate a new request. Thus the system is provided with what documents the simulated user interacted with including data like dwell times, relevance feedback, etc. The process would continue until some stopping criteria is met (i.e. x relevant documents are found, y queries are submitted, etc). Here, an agent is defined that specifies / instantiates a simulated user, which is designed to interact with a pre-defined system API. This approach assumes that relevance judgments already exist. This is because to make the approach feasible, pre-existing judgements are needed for the simulated users so that they can provide feedback to the systems. While this kind of simulation does not allow the full the spectrum of unbounded possibilities because it is constrained by the set of pre-existing judgements available for a given topic, it does provide significant more evaluation possibilities over sessions. Implementing such a solution within a common evaluation forum though introduces additional complexities in terms of infrastructure required by the evaluation forum and conformance to the infrastructure by participants. However, these kinds of evaluations are growing in popularity and a number of examples of this type of simulation have been produced recently.

Unbounded and Unlimited Interaction:
This is perhaps the holy grail of simulation where it will be possible to provide a complex information need and task to a simulated user and let it engage with an IR system. Having such simulated users would be on par with developing a Turing test of Interactive Information Retrieval. Such simulated users to be effective would need to have the capacity to interpret and judge the information that they are encountering and to handle the complexities and nuances of the interfaces that they are presented. This provides an interesting and challenging area of future research that will open up many interesting sub-problems and new tasks in designing a simulated and "intelligent" user.

In each of the above approaches a simulated user is instantiated and uses an IR system, for a given task/ scenario. A goal is defined in terms of some objective measure, for example: total amount of gain in a given amount of time or interaction, recall given a certain amount of interaction, etc. In this unbounded and unlimited approach, goals may even be undefined, poorly specified, or focused on other evaluation measures (like fun, engagement, etc). This brings up an important point, in situations, such as searching for "lol cats", searching for pornography, etc, where the end goal is not as clear, or it is very clear but different to measure, where there are externalities involved (like other pressing tasks, interruptions, multiple searchers, etc), it is difficult to simulate. Largely this is because the tasks are vaguely defined and under researched, whereas for standard ad hoc retrieval, for instance, we have pre-existing test collections and much research has been undertaken examining the situation.

What-If Simulations of Interaction:

Building simulations can be used to evaluate systems as well as to evaluate user behaviors. So rather than trying to create simulated users that replicate actual user behaviors, instead, we can create simulated users that act in different ways. This is a powerful type of simulation which enables us to explore a range of interaction possibilities, strategies, and information behaviors. Here, simulated behaviors can be created and compared to evaluate how well a user could use an IR system over a session for a given task. This may lead to finding out new and improved behaviors or finding that particular behaviors are more appropriate for different systems. For instance, we could simulate a user with 4 arms and a tail with opposable thumbs, and evaluate how well they would perform on a given system. This approach has the advantage that we are not constrained by reality and affords the researchers more luxuries in speculating and evaluating users and systems.

Future collaboration with host institution (if applicable);

Future collaboration with NII will take place through the Text REtrieval Conference (TREC) working closely at better constructing the evaluation framework to be used.

Projected publications / articles resulting or to result from the grant (ESF must be acknowledged in

publications resulting from the grantee's work in relation with the grant);

The intention is to (a) publish the scientific report produced out of this meeting at SIGIR Forum and (b) push forward some of the ideas produced in this meeting to the TREC Session Track which I will be coordinating for a fourth consecutive year.