EVALUATION

Nick Belkin, Pia Borlund, Ben Carterette, Diane Kelly, Bill Kules, and Mark Smucker

Our group considered the problem of evaluation. This area has been extensively examined, and we did not attempt to enumerate the variety of definitions and challenges that have been proposed. Instead, we focused on a few themes that seem common throughout the literature. Traditionally, we have considered IR as a tool for finding documents, and evaluation consisted of assessing how good the documents are. This approach has been questioned and alternative models and measures have been proposed. For example, in addition to evaluating documents, we can evaluate both the information seeking and retrieval process, as well as the end results. We can consider evaluation at multiple levels, too. The scope of an evaluation can range (at least) from the entire session (with respect to the motivating task), to the usefulness of a particular system support (with respect to its own design intention or the higher level task). The evaluation can also be framed in terms of how well the system (or component) supports the person in optimizing the search process.

We identified several evaluation challenges as particularly apropos for task-based search, including ecological validity, trustworthiness of results and generalizability, domain-specific search. Creating measures that can be used within and across studies of users is an important challenge. Given a task-type, it is highly desirable to have a recommended set of measures that have been validated and calibrated so that they can be used to compare results between studies. This could be represented as a matrix of task-type by recommended measure. Other challenges include:

- Better ways to get people to opt into exploratory research of usefulness
- What constitutes a realistic environment in the lab
- How to minimize demand effects
- What data is needed to support simulation evaluation – build models & measures - Characterizing the space of ways to get to an outcome

Usefulness of Task-Based Searches

Nick Belkin, Ben Carterette, and Mark Smucker

Problem / Motivation / Goals

In traditional information retrieval evaluation frameworks, we consider documents to be relevant or non-relevant and a user's interaction with a search system is reduced to being the rate and amount of relevant documents retrieved by a user. In task-based search, we do not know how users value their usage of search engines in support of their tasks. For example, we would like to, but do not know the answers to the following questions:

- Are there common stages that users experience while working on tasks?
- What are the intentions of users at these various stages of their task?
- How do users measure / describe the usefulness of their experience in the various stages (if stages exist, otherwise steps)?
- How do users measure / describe the overall usefulness of their interaction with the search engine?

Without the basic knowledge of how users value task-based searches, we are unable to properly evaluate retrieval systems' support for task-based search tasks.

Approaches

We propose two approaches to help us better understand how users value search systems for task-based search tasks. The first is a broad approach that would ask study participants to recall tasks and then discuss the usefulness of their actions. We envision using extensive instrumentation to collect detailed usage data from participants, e.g. eye-tracking laptops. The study would be conducted over a long time period of at least 3 months with weekly interviews conducted. In addition to instrumented laptops, we would ask participants to maintain diaries of their tasks and also to have the user flag relevant activity for later review with the researchers. The idea behind this broad approach is that recorded behavioral data often does not reveal users' motivations and actual tasks.

The second approach would be a controlled lab study. In this study, we would have assigned tasks that differ in the dimensions of expected behavior. Again, we would use a highly instrumented system. By using stimulated recall with the participants, we could carefully measure micro-actions without interruption of the activity. This second approach would logically follow after the conclusion of the first.

Nature of the Interview

The "broad approach" above calls for participants to use a highly instrumented laptop, but even with a collection of data, we still need some way to identify episodes and critical incidents to review with the participants. Some possible aids that could be used to identify task-based searches include:

- Elicitation question: This is a question designed to help participants recall useful incidents to examine with the researcher.
- Various aids such as diary entries, flagged points in time, and replayed screenshots / video: Each of these involve either the participant making note of interesting events when they occur or helping the participant recall the event from a replay of activity.

Likewise, the researcher could attempt to identify task-based searches from the record data, and it would be good to compare participant identified vs. experimenter identified episodes.

Once a task-based search episode is identified, the researcher would likely proceed through a series of questions with the participant including:

- What was your intention with this search?
- Is this task a one-off or a regular task?
- This looks where you started, is this right or was it earlier?

For the various events occurring in an episode, the questions would include:

- Can you remember what you wanted to do?
- Do you feel you were successful here?
- What makes you feel that way?
- How do feel about this search in the whole?

Once the quantitative behavior data and qualitative interview data has been collected, it would be necessary to analyze the interview content and tie it to the behavior data.

Analysis of Data

The interviews would produce a large amount of verbal data. In this data, the hope is that the participants would use words that are indicative of measures of usefulness. Likewise, we anticipate that these searches will go through various stages that can be generalized to most searches. With

stages identified, we would then look for sequences of behavior that could be used to identify a stage. Likewise, we would attempt to classify the various identified tasks according to their characteristics.

Challenges / Resources Required / Caveats

We have outlined a large study involving the deployment of expensive laptops. Assuming the technology works, in a short period of time a large amount of data will be produced, and the amount of data may overwhelm attempts to understand it. Likewise, the verbal data is less reliable measure of value than we would like.

Impact

Today evaluation of search is tied to a notion that documents deliver the only value to the user. We intuitively know that there is value to search that goes beyond a set of relevant documents. By focusing on what users find useful, we have the potential for improving the evaluation of retrieval systems, and as our ability to measure effectiveness improves, so does our ability to build better systems.

Plan for future

As outline above, our proposal is to conduct an exploratory study to identify intentions and measures of usefulness in task-based search. Following this study, we will need a controlled experimental study to test the hypotheses generated by this study. The controlled study will try to determine if, for given tasks, do users have the same intentions and notions of usefulness that we identified in the naturalistic study? Assuming success with the exploratory and lab studies, we would then work on creation and validation of explicit measures for support of motivating task types and search intentions.

Meta-framework for comparable task-based evaluation

Pia Borlund, Diane Kelly, Bill Kules

On day 2, a small subgroup examined the challenge of comparability of evaluations. The current variety of research and reporting practices makes integration of multiple studies difficult. This is holding back meta-evaluations. Some modest changes in research practices could increase the comparability and interoperability of studies, through the clear and consistent description of studies and their results. As a start, we identified a meta-framework for studies that includes six elements: tasks, study design, measures, analysis, reporting and a matrix of measures. For each element, we explored actions that could be taken to document best practices for the design of each element of the evaluation. As shown in the table, reporting practices for users, tasks, etc. are well-established. Practices for tasks, study design and analysis methods could be synthesized from existing literature. The design of measures and a matrix for recommending which measures to use by task-type will require new research.

Elements	Actions
Tasks	Synthesis (being done)
Study Design	Synthesis (to-do)
Measures	R&D Needed
Analysis	Synthesis (to-do)
Reporting practices • Users • Tasks • …	Now
Matrix	Future

The potential impacts of such a framework include facilitation of future meta-studies, as well as longer-term historical analysis (e.g. 50+ years). A framework will also encourage better reflection on specific practices and their limitations. It could also yield educational benefits to researchers and students and enhance research integrity. Unintended consequences could include mimicry (that is, encouraging copying rather than learning) and overall rigidity that slows innovation. Researchers may also be resistant to change their practices.

Nevertheless, such a framework could be beneficial. Several possible steps could be taken, starting with proposals for best practices on reporting tasks. Design and analysis elements could be synthesized from the existing literature. More research and development will be required for the analysis element. The entire framework will need to be evaluated. This could be initially accomplished by applying it to analyze existing studies.