Usefulness as the Evaluation Criterion for Interactive Information Retrieval

Nicholas J. Belkin
School of Communication & Information
Rutgers University
belkin@rutgers.edu

The Goal of Information Retrieval

- People engage in information seeking in order to accomplish some task, or achieve some goal
  - Because they find themselves in a problematic situation
- IR systems should therefore be concerned with supporting people in accomplishing the tasks or achieving the goals, which led them to engage in information seeking
  - Or, perhaps, helping people to resolve their problematic situations

Operationalizing the IR Goal

- How can an IR system provide such support?
  - What is it feasible/reasonable to expect of an IR system?
  - To what extent can a system directly support accomplishment of a task, achievement of a goal, resolution of a problematic situation?
  - The answers establishe a criterion or criteria for evaluation
- How can the degree to which an IR system has provided such support be measured?
  - Depends on criterion/criteria
  - What can actually be measured, and how?

Some IR History

- Earliest operationalizations of the goal of IR
- Kent, Moers, Cleverdon (late 1950s, early 1960s), later, everyone else
  - Providing **relevant** documents to the person
- What does "relevant" mean?
  - On the topic? Something else?
- What does it mean to "provide"?
  - In response to a request? Something else?

A Hierarchy of IR System Goals

- Task accomplishment– too difficult to measure, more than is reasonable to expect of IR system
- Resolution of problematic situation necessary for accomplishing task, difficult to describe and measure
- Resolution of ASK necessary for resolution of problematic situation, requires changing state of knowledge, possible to describe and measure, but highly person/context specific
- Provision of useful documents is a means for resolution of ASK (changing state of knowledge) but many factors impact usefulness of documents
- For a document to be useful, it must be **relevant**

Provision of Relevant Documents

- A person describes what it takes for a document to be relevant (i.e. provides a "query")
- The IR system interprets that description, and
- Returns to the person a set or list of documents predicted to be relevant to the query

Measuring IR System Effectiveness

- According to a user model
  - What the person requires of the system in terms of relevant documents
  - Historical model: all and only relevant documents
- Abstract effectiveness
  - Ability to find relevant documents
  - Ability to discriminate between relevant and non-relevant documents
- Measured with respect to *a* response to *the* query

The Cranfield/TREC Evaluation Model

- Relevance as the evaluation criterion
- Comparison between systems as the means for evaluation
- Single response to a single query as the item to be measured
- Recall and precision as the measures of performance
  - Combination of abstract effectiveness and a specific user model

Problems with this Model

Has served us well in many respects, but

- People engage in information seeking episodes
- ASK hypothesis suggests dynamic information seeking, as user and system change
- People have different intentions during an information seeking episode
- Effectiveness of support for an information seeking episode cannot be evaluated solely by provision of relevant documents

Upshot

- Cranfield/TREC evaluation model cannot be extended to evaluate IR systems in their support of interactive information retrieval
  - Different intentions during an information seeking episode
  - Changing state of the person during the information seeking episode
  - Differences between people
- Consider, e.g., the TREC Interactive Track

Interactive Information Retrieval

- An information seeking episode is instigated by a motivating task or goal, and is an attempt to resolve a problematic situation
- An information seeking episode consists of a sequence of a person's interactions with information objects and the IR system
- These interactions may each be undertaken with different intentions
- An information seeking episode concludes (ideally) with resolution of the problematic situation
- Resolution of the problematic situation results (ideally) in accomplishment of the task or goal

Possible Evaluation Paradigms for Interactive Information Retrieval

- Satisfaction with a search
  - Multi-dimensional, highly subjective, difficult to interpret and compare (Tagliacozzo, 1977)
  - Evaluation of search results as a whole a possible measure (Su, 1998)
- Usability of system
  - Not easy to relate to eventual outcome of information seeking episode
  - Effort required to achieve goal a possible measure

Usefulness as an Evaluation Criterion

- Usefulness as a category can encompass many other criteria, depending upon the goal according to which it is applied
- Usefulness can be applied to evaluate not just support for an entire information seeking episode, but also to evaluate support for the separate intentions within an information seeking episode
- Usefulness could be applied to evaluate the sequence of types of support offered the searcher during the information seeking episode

Three Levels of Usefulness Evaluation

- How useful was the system in helping to accomplish the motivating task/goal
- How useful was the system support in achieving the goal of each search intention
- How useful was the achievement of each search intention to the accomplishment of the task/goal

Measuring Usefulness

- Characterizing the motivating task/goal
  - Elicitation or control of task/goal, mapped to a taxonomy
- For search session as a whole
  - Eliciting usefulness judgment
  - Measuring task/goal accomplishment
- For each search intention *qua* intention
  - Eliciting usefulness or accomplishment judgment
  - Measuring according to specific criteria for intention
- For each search intention, wrt overall goal
  - Eliciting usefulness of accomplishment of intention
  - Comparing alternative sequences of intentions

Some Problems with Usefulness & Some Possible Solutions

- Establishing methods that allow replicability and comparability
  - Consistent user models
  - Controlled task types
  - Possible simulation
- Eliciting usefulness judgments
  - At intention level, use only "objective" measures
  - At search session level, rely on measures of task performance
- Scale
  - Simulation?

An Attempt at Applying Usefulness

Two interrelated projects which attempt to

- Identify coherent sequences of behaviors during an information seeking episode
- Relate behaviors to specific, different search intentions during an information seeking episode
- Identify criteria for support for different search intentions
- Test measures and methods for evaluation of support for different intentions
- Test measures and methods for evaluation of support for controlled motivating task types

So Far

- Identified coherent sequences of eye fixation-related behaviors
  - Existing data from searches in four different task types, varied according to three different task facets
  - Around 8 – 10 classes of behaviors
- Different task types / facet values associated with different sequences of behavior classes

What's Next (1)

- Stud[y|ies] of searching with controlled motivating tasks; logging behaviors and eliciting:
  - Intentions wrt to each query interval;
  - Success in achieving intentions;
  - Success in accomplishing task;
  - Usefulness of system support for accomplishing task.

What's Next (2)

Based on these studies

- Identifying search intentions and criteria for evaluation of their accomplishment
- Identifying relationships of behaviors to search intentions
- Identifying "optimal" sequences of intentions/ behaviors
- Characterizing usefulness measures for accomplishing different task types

What's Next (3)

- New sets of studies, in lab and in situ, testing
  - Measures of usefulness of support for different search intentions, by comparing "objective" measures based on intention-specific criteria to usefulness judgments of those support techniques
  - Methods for collecting data required for these measures
  - Measures of usefulness of support for task accomplishment, by comparing "objective" measures of task performance to usefulness judgments of search system support

In (Only Partial) Conclusion

- Thanks for listening
- Thanks for comments, suggestions, criticisms
- Hoping to address at least some of the problems associated with usefulness evaluation while at GESIS
- Thanks to Google for a Faculty Research Award, and Michael Cole for doing the work
- Thanks to NSF for Grant No. IIS-1423239
- And, for the future we'll probably abide by …

Das kölsche Grundgesetz,

- Artikel 1: „Et es wie et es.“
- Artikel 2: „Et kütt wie et kütt.“
- Artikel 3: „Et hät noch immer joot jejange.“
- Artikel 6: „Kenne mer nit, bruche mer nit, fott domet.“ (But, we hope to know (learn) a lot!)