A Model for Evaluation of Interactive Information Retrieval

Nicholas J. Belkin, Michael Cole, and Jingjing Liu
School of Communication & Information
Rutgers University
4 Huntington Street, New Brunswick, NJ 08901, USA
{belkin, m.cole}@rutgers.edu, jingjing@eden.rutgers.edu

Categories and Subject Descriptors
H.1.2 [**Models and Principles**]: User/Machine Systems – *human information processing* H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process*

General Terms
Measurement, Performance, Experimentation, Human Factors

Keywords
Evaluation, Information seeking, Interaction, Usefulness

1.INTRODUCTION
Research in information retrieval (IR) has expanded to take a broader perspective of the information seeking process to explicitly include users, tasks, and contexts in a dynamic setting rather than treating information search as static or as a sequence of unrelated events. The traditional Cranfield/TREC IR system evaluation paradigm, using document relevance as a criterion, and evaluating single search results, is not appropriate for many circumstances considered in current research. Several alternatives to relevance have been proposed, including utility, and satisfaction. We suggest an evaluation model and methodology grounded in the nature of information seeking and centered on *usefulness*. We believe this model has broad applicability in current IR research.

2.INFORMATION SEEKING
As phenomenological sociologists note, people have their life-plans and their knowledge accumulates during the process of accomplishing their plans (or achieving their goals). When personal knowledge is insufficient to deal with a new experience, or to achieve a particular goal, a *problematic situation* arises for the individual and they seek information to resolve the problem [1]. Simply put, information seeking takes place in the circumstance of having some goal to achieve or task to complete.

We can then think of IR as an information seeking episode consisting of a sequence of interactions between the user and information object(s) [2]. Each interaction has an immediate goal, as well as a goal with respect to accomplishing the overall goal/task. Each interaction can itself be construed as a sequence of specific *information seeking strategies* (ISSs) [3].

We believe appropriate evaluation criteria for IR systems are determined by the system goal. The goal of IR systems is to support users in accomplishing the task/achieving the goal that led them to engage in information seeking. Therefore, IR evaluation should be modeled under the goal of information seeking and should measure a system's performance in fulfilling users' goals through its support of information seeking.

3.GOAL, TASK, SUB-GOAL & ISS
In accomplishing the general work task and achieving the general goal, a person engaged in information seeking goes through a sequence of information interactions (which are sub-tasks), each having its own short term goal that contributes to achieving the general goal. Figure 1 illustrates the relationships between the task/goal, sub-task/goal, information interaction, and an ISS.

Let us give an example. Suppose someone in need of a hybrid car wants to choose several car models as candidates for further inspection at local dealers. The *problematic situation* [1] here is that he lacks knowledge on hybrid cars. His general *work task* is seeking hybrid car information and deciding which models he should look at. He may go through a sequence of steps which have their own *short-term goals*: 1) locating hybrid car information, 2) learning hybrid car information, 3) comparing several car models, and 4) deciding which local dealers to visit. In each *information interaction* that has a short-term goal, he may go through a sequence of *ISSs*. For example, searching for hybrid car information can consist of querying, receiving search results, evaluating search results, and saving some of them.

There are several general comments about Figure 1. First, it shows only the simplest linear relations between the steps along the time line. In fact, the sequence of steps/sub-goals/ISSs could be non-linear. For instance, on the sub-goal level, after learning hybrid car information, the user may go back to an interaction of searching for more information. Another example on the ISS level is, after receiving search results, the user may go back to the querying step.

Second, the contribution of each sub-goal to the general goal may change over time. For instance, suppose in one information interaction, the user looks at information of car model 1 and decides to choose it as a final candidate. After he learns about car model 2, which outperforms car model 1 in all aspects, he removes model 1 from the candidate list. Therefore, some steps in the sequence (choosing car model 1) may contribute to the sub-goal positively, but it contributes to the final and overall goal negatively in that car model 1 is eventually removed.

Third, the leading goal of this task is, or can be taken to be, relatively stable over the course of the interaction. Different users can and will do different things to achieve similar leading goals. Some of differences in these sequences may be characteristics of classes of users, for example, high/low domain knowledge, cognitive capacities, and of task types, including task complexity.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

4.AN EVALUATION MODEL

We suggest IR evaluation should be conducted on three levels. First, it should evaluate the information seeking episode as a whole with respect to the accomplishment of the user's task/goal. Second, it should assess each interaction with respect to its contribution to the accomplishment of the overall goal/task. Third, it should assess each interaction, and each ISS, with respect to its specific goal. In this framework, an ideal system will support the task accomplishment by presenting resources and user support in an optimally-ordered minimum number of interaction steps.

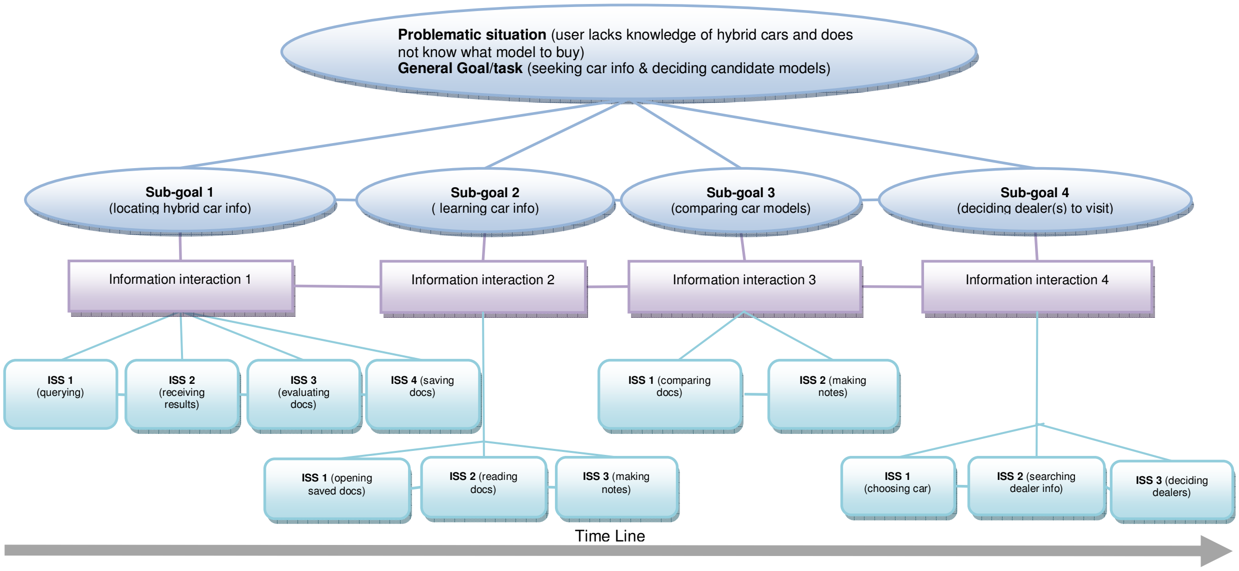
4.1Criterion: Usefulness

We suggest that *usefulness* is an appropriate criterion for IR evaluation. Usefulness should be applied both for the entire episode against the leading (work) task/goal and, independently, for each sub-task/interaction in the episode. Specifically, 1) How useful is the information seeking episode in accomplishing the leading task/goal? 2) How useful is each interaction in helping accomplish the leading task? 3) How well was the goal of the specific interaction accomplished? From the system perspective, evaluation should focus on: 1) How well does the system support the accomplishment of the overall task/goal? 2) How well does the system support the contribution of each interaction towards the achievement of the overall goal? 3) How well does the system support each interaction?

4.2Measurement

Operationalization of the criterion of usefulness will be specific to the user's task/goal, at the level of the IR episode; to the empirical relationship between each interaction and the search outcome, at the level of contribution to the outcome; and to the goals of each interaction/ISS at the third level.

Examples at each level might be: the perceived usefulness of the located documents in helping accomplish the whole task; task accomplishment itself; the extent to which documents seen in an interaction are used in the solution; the degree to which useful documents appear at the top of a results list; and the extent to which suggested query terms are used, and are useful. Identifying specific measures and how to achieve them are clearly difficult problems. However, we believe evaluation of IR systems should be grounded in the nature of the information seeking process that is the *raison d'etre* for these systems. Comments are welcome.


Problematic situation (user lacks knowledge of hybrid cars and does not know what model to buy)
General Goal/task (seeking car info & deciding candidate models)
Sub-goal 1 (locating hybrid car info)
Sub-goal 2 (learning car info)
Sub-goal 3 (comparing car models)
Sub-goal 4 (deciding dealer(s) to visit)
Information interaction 1
Information interaction 2
Information interaction 3
Information interaction 4
ISS 1 (querying)
ISS 2 (receiving results)
ISS 3 (evaluating docs)
ISS 4 (saving docs)
ISS 1 (opening saved docs)
ISS 2 (reading docs)
ISS 3 (making notes)
ISS 1 (comparing docs)
ISS 2 (making notes)
ISS 1 (choosing car)
ISS 2 (searching dealer info)
ISS 3 (deciding dealers)
Time Line

Evaluation based on the following three levels:

1. The usefulness of the entire information seeking episode with respect to accomplishment of the leading task;
2. The usefulness of each interaction with respect to its contribution to the accomplishment of the leading task;
3. The usefulness of system support toward the goal(s) of each interaction, and of each ISS.

Figure 1. An IR Evaluation Model

5.ACKNOWLEDGMENTS

These ideas have benefited from discussion at the 2009 Dagstuhl Seminar on Interactive Information Retrieval, especially the contributions of Pertti Vakkari, Kal Järvelin, and Norbert Fuhr.

This work is supported by IMLS grant LG-06-07-0105-07.

6.REFERENCES

[1] Schutz, A. & Luckmann, T. (1973). *The structures of the life-world.* Evanston, IL: Northwestern University Press.

[2] Fuhr, N. (2008). A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11: 251-265.

[3] Yuan, X.-J. & Belkin, N.J. (2007). Supporting multiple information-seeking strategies in a single system framework. In *SIGIR '08* (pp. 247-254). New York: ACM.