Usefulness as the Criterion for Evaluation of Interactive Information Retrieval

M. Cole, J. Liu, N. J. Belkin, R. Bierig, J. Gwizdka, C. Liu, J. Zhang, X. Zhang
School of Communication and Information
Rutgers University
4 Huntington Street, New Brunswick, NJ 08901, USA
{m.cole, belkin, bierig, jacekg}@rutgers.edu, {jingjing, changl, zhangj}@eden.rutgers.edu, xiangminz@gmail.com

ABSTRACT
The purpose of an information retrieval (IR) system is to help users accomplish a task. IR system evaluation should consider both task success and the value of support given over the entire information seeking episode. Relevance-based measurements fail to address these requirements. In this paper, *usefulness* is proposed as a basis for IR evaluation.

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 interactive information retrieval (IIR). Several alternatives to relevance have been proposed, including utility and satisfaction. We have suggested an evaluation model and methodology grounded in the nature of information seeking and centered on *usefulness* [1] [2]. We believe this model has broad applicability in current IR research. This paper extends and elaborates the model to provide grounding for practical implementation.

2 INFORMATION SEEKING
As phenomenological sociologists (e.g., [7]) 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 [7]. 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 objects [4]. 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) [8].

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.

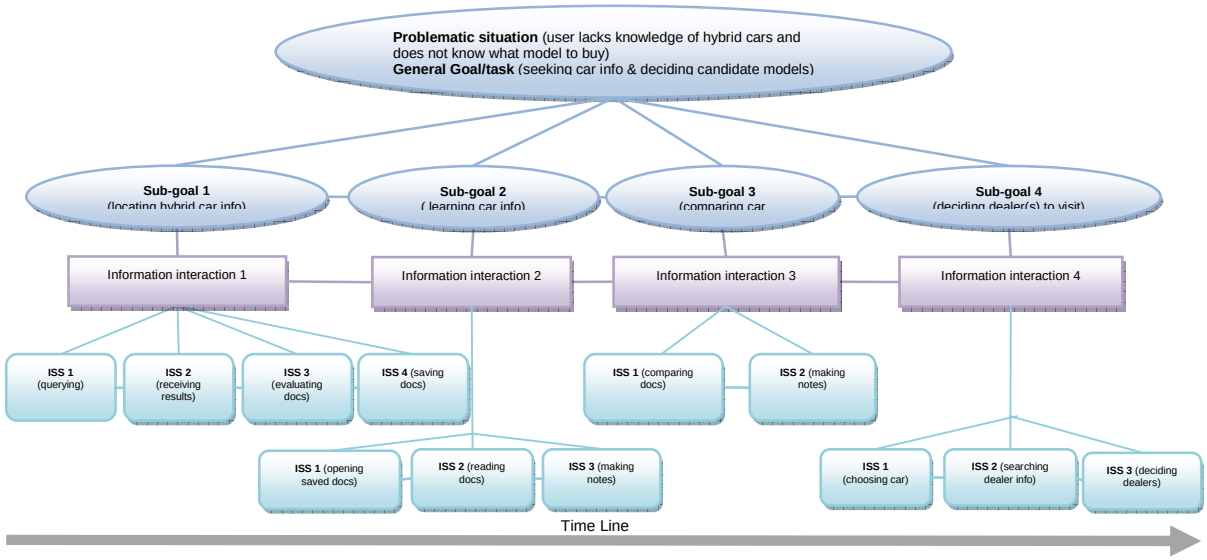
3 GOAL, TASK, SUB-GOAL & ISS
In accomplishing the general work task and achieving the general goal, a person engaged in information seeking goes through a sequence of information interactions (which are sub-tasks), each having its own short term goal that contributes to achieving the general goal. Figure 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* [7] is that he lacks knowledge on hybrid cars. His general *work task* is seeking hybrid car information and deciding at which models he should look. 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* with a short-term goal, he may go through a sequence of *ISSs*. For example, searching for hybrid car information may consist of querying, receiving search results, evaluating search results, and saving some of them.

There are several general comments. First, Figure 1 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. At the ISS level, 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

HCIR Workshop 2009, Washington, D.C., October 23, 2009.



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 IIR Evaluation Model

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 the 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.

4 AN EVALUATION MODEL

Fundamentally, we are interested in why a person engages an information need and how an interaction session contributes to meeting that need. It follows one must provide a measurement for the session as a whole and for the session constituents.

4.1 Three levels of evaluation

A user makes progress towards a goal by virtue of the results of interactions with the system. Support of results and process are two aspects of system performance. Evaluation of a system should center on how well the user is able to achieve their goal, the process of helping the user identify and engage in appropriate interactions, and the relationship of the results of those interactions to the progress toward and accomplishment of the goal. IR evaluation should then 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, meaning explicitly the effectiveness of support for each ISS, with respect to its immediate goal. Third, it should assess each interaction with respect to its contribution to the accomplishment of the overall task/goal.

An ideal system will support its users' task accomplishment by presenting resources and user support in an optimally-ordered minimum number of interaction steps (cf. [3]). Resources and user support should address not only search result content, i.e., techniques to rank the most relevant documents at the top, but they should also be manifest in the system interface, including search interface, result display, and various ways to support general task accomplishments. For example, the system could have a function of comparing pages that users have seen for them to better understand or summarize what they have learned about the task topic, so as to help them in solving the task. As another example, the system may have a place for the users to make notes, or create document drafts, which on one hand, is a way of helping users start generating their task-solving documents, and, on the other hand, are helpful for relevance feedback/query reformulation.

4.2 Criterion: Usefulness

We propose *usefulness* as the criterion for IIR evaluation. Existing measures of IR performance are inadequate for the proposed IIR evaluation model.

The sense of usefulness we have in mind is more general than relevance, which has come, for historical reasons [1], to be the received basis for measuring IR systems. Like relevance, people are able to give usefulness judgments as intuitive assessments that do not turn on understanding a technical definition. Usefulness, however, is suited to interaction measurements in ways relevance-based systems cannot address.

The problem of measuring IIR has recently received attention in terms of formal models (e.g. [4]) and the relation of local interactions to realization of search session outcomes (e.g. [5]). Usefulness measurements are distinguished from session-level measurements like Järvelin, et al.'s session-based discounted

cumulative gain (sDCG) [5] in that usefulness explicitly considers the session as a whole. sDCG does not support judgment of relevance to the whole session or how results from an interaction step might be integrated into the whole. It depends on the assumption the only thing that matters is the relevance of the local interaction and the incremental change it makes on the history of relevance judgments to that point.

Usefulness is specifically distinguishable from relevance in several dimensions. Most strikingly, a usefulness judgment can be explicitly related to the perceived contribution of the judged object or process to progress towards satisfying the leading goal or a goal on the way. In contrast to relevance, a judgment of usefulness can be made of a result or a process, rather than only to the content of an information object. It also applies to all scales of an interaction. Usefulness can be applied to a specific result, to interaction subsequences, and to the session as a whole. Usefulness, then, is more general than relevance, and well-suited to the object of providing a measurement appropriate to the concept of task goal realization.

This does not deny the importance of relevance as a specific measurement to be used in appropriate circumstances to determine usefulness. For example, relevance can be used as a usefulness criterion for interaction steps where the immediate goal is to gather topical documents. Here, it is the aboutness of a document that constitutes its usefulness to advancing the task, so relevance is the appropriate usefulness criterion. This example illustrates a larger point tied with the generality of usefulness as a measure. Measuring usefulness relies on adopting appropriate and varied criteria, even within a task session. Examples of such criteria include explicit judgments including relevance and usefulness, and implicit markers, such as decision and dwell times on documents, number of steps to complete a sub-goal, user's actions to save, revisit, classify and use documents, and issue and reformulate queries. Researchers already use specific criteria such as these for evaluation. One consequence of adopting usefulness is that several measures should be used, and perhaps only for specific segments in the episode. Identifying which measures are important for episode components and for the entire episode must be experimentally determined.

Usefulness should be applied both for the entire episode against the leading goal/task 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?

4.3 Measurements

Identifying specific measures of usefulness and how to obtain them are clearly difficult problems. The most important aspect of this evaluation framework is that it depends crucially upon specification of a leading task or goal whose accomplishment can itself be measured.

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

Examples at each level might be: the perceived usefulness of the located documents in helping accomplish the whole task; task accomplishment itself, in terms of correctness, effort, or time; the extent to which systems suggestions as to what to do are taken up; 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.

As an example, consider the hybrid car information seeking episode and focus on just the leading goal/task, sub-goal 1 and the information interaction 1 with its four ISSs. To demonstrate how the criterion of usefulness can be operationalized, the evaluation could be approached from the following aspects (this is not intended as an exclusive list):

- **at the level of the whole episode [leading goal/task]**
  - *accomplishment of the task [result]*
    - How well did the user successfully select candidate car models? [correctness]
    - How many steps (e.g., interactions, ISSs) did the user go through for the whole task? [effort]
    - How long did the user spend to complete the whole task? [time]
  - *support to the information seeking episode [process]*
    - How useful was the system in supporting identification of appropriate sub-goals in selecting hybrid car models?
    - Were system suggestions on what to do (e.g., a system suggesting four task steps: locating information, learning, comparing, and deciding) accepted?
    - How well did the system support the user in choosing an appropriate sub-task sequence?
- **at the level of the information interaction/sub-goal [sub-goal 1/information interaction 1]**
  - *accomplishment of the sub-goal [result]*
    - How well did the user successfully locate hybrid car information? [correctness]
    - How many steps (e.g., ISSs) did the user go through in locating car information? [effort]
    - How long did the user spend to locate car information? [time]
  - *support to information interaction 1[process]*
    - How useful was the system in supporting users to identify appropriate ISSs in locating hybrid car information?
    - Were system suggestions on what to do (e.g., suggesting a user should now query, view results, evaluate results or save documents) accepted?
    - How well did the system support the user in choosing an appropriate ISS sequence?
- **at the level of the contribution of the sub-goal to the leading goal [sub-goal 1 to the leading goal]**
  - *accomplishment of the contribution [result]*
    - How much did locating car information contribute to the whole task of selecting candidate car models?
  - *support to this contribution [process]*
    - How useful was the system in supporting users to locate car information in order to finally select candidate car models?

- **at the level of the ISSs [ISSs 1-4 information interaction 1]**
  - How useful were suggested queries/terms for formulating queries? [ISS1]
  - How much were the suggested queries/terms used? [ISS1]
  - How well does the system support evaluation of retrieved documents? [ISS3]
  - How well does the system support saving or retaining the retrieved, or useful, documents? [ISS4]
- **at the level of the contribution of each ISS to the sub-goal or leading goal [ISSs 1-4 to sub-goal 1 and leading goal]**
  - How useful were the suggested queries/terms (for systems with query formulation assistance) for locating car information? [ISS1 to sub-goal 1]
  - How well did the system rank documents? (using relevance and various other measures: precision, DCG, etc.) [ISS2 to sub-goal 1]
  - How useful was each viewed document in helping users locate hybrid car information? [ISS2 to sub-goal 1]
  - How useful was each viewed document in helping users select the candidate car models? [ISS2 to leading goal]

4.4 Experimental frameworks for IIR system measurement

One challenge in measuring the performance of IIR systems is to move beyond the Cranfield and TREC relevance-based models. Several experimental frameworks are available to measure system performance over interactive sessions.

Traditional user-studies can be used by setting a task with a measurable outcome that is related to information seeking activities. Systems are then compared by both outcome and the interaction path taken to task completion. Our proposal addresses how the interaction path can be assessed to measure its contribution to the outcome.

The limitations of user-studies are scale-related. One can address only a small number of tasks with a limited number of subjects. User-studies have the virtue of well-specified tasks and the ability to collect many details about users and their interactions.

An alternative framework, in the spirit of A-B system comparisons often used in commercial settings, is to make available two versions of a system and compare measures as people make use of the system (e.g. [6]). A big advantage of this approach is the ability to conduct large-scale tests with many users and (implicitly) many tasks. The limitation is that one needs to infer properties of the tasks and also the usefulness of the system response to meeting the needs of the users. One difficult technical issue is the identification of sessions to enable session-level results analysis.

A third, somewhat intermediate, approach to achieve reasonable scale with enough detail to enable a rich assessment of system performance for user task support, is to build a reference database of session interactions. This might be assembled in a cooperative effort and made available to research groups to generate system performance results. Such a usefulness-based interaction database would presumably include user models to choose interaction outcomes depending on the choices offered by the system and the support provided by the system at each step along the way. Such a database might be generated from uniformly-instrumented user studies and a reference user model(s).

5 CONCLUSION

Information retrieval is an inherently and unavoidably interactive process, which takes place when a person faces a problematic situation with respect to some goal or task. Thus, evaluating IR systems must mean both evaluating their support with respect to task accomplishment, and evaluating them with respect to the entire information seeking episode. Past, and most current approaches to IR evaluation, as exemplified by TREC, fail to address either of these desiderata, focusing as they do on relevance as the fundamental criterion, and on effectiveness of system response to a single query. In this paper, we propose an alternative evaluation model which attempts to address both of these issues, based on the criterion of *usefulness* as the basis for IR evaluation. Although our proposed model clearly needs more detailed explication, we believe that it offers a useful basis from which realistic and effective measures and methods of IR evaluation can be developed.

6 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, An earlier version of this paper was presented at the SIGIR 2009 Workshop on The Future of IR Evaluation, and discussion there has substantially influenced this version. This work is supported by IMLS grant LG-06-07-0105-07.

7 REFERENCES

[1] Belkin, N.J, Cole, M., and Bierig, R. (2008). Is relevance the right criterion for evaluating interactive information retrieval? In *Proceedings of the SIGIR 2008 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level judgments*, (Singapore, 2008). Retrieved from: http://research.microsoft.com/en-us/um/people/pauben/bbr-workshop/talks/belkin-bbr-sigir08.pdf on August 24, 2009.

[2] Belkin, N.J., Cole, M. and Liu, J. (2009). A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation* (Boston). IR Publications, Amsterdam. 7-8.

[3] Belkin, N.J., Cool, C., Stein, A. and Thiel, U. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(30). 379-395.

[4] Fuhr, N. (2008). A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11.251-265.

[5] Järvelin, K., Price, S.L., Delcambre, L.M.L., and Nielsen, M.L. (2008). Discounted cumulative gain based evaluation of multiple-query IR sessions. In *Proceedings of the 30th European Conference on Information Retrieval* (Glascow, Scotland, 2008), Springer-Verlag. 4-15.

[6] Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Edmonton, Canada, 2002) ACM. 133-142.

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

[8] Yuan, X.-J. and Belkin, N.J. (2008). Supporting multiple information-seeking strategies in a single system framework. In *Proceedings of the 31st ACM SIGIR International Conference on Research and Development in Information Retrieval* (Singapore, 2008). ACM. 247-254.