



EVALUATION CHALLENGES AND DIRECTIONS FOR INFORMATION- SEEKING SUPPORT SYSTEMS

Diane Kelly, *University of North Carolina at Chapel Hill*

Susan Dumais, *Microsoft Research*

Jan O. Pedersen, *A9.com*

ISSSs provide an exciting opportunity to extend previous information-seeking and interactive IR evaluation models and create a research community that embraces diverse methods and broader participation.

You're thinking about booking a vacation and would like to go someplace new. What process would you follow to gather candidate destinations and what systems, if any, would you consult? Would you regard this as a search task or something more general? What would constitute a successful outcome?

Search tools that support these sorts of open-ended tasks are referred to as information-seeking support systems. Central to the development of ISSSs is the question of evaluation. What does it mean for an ISSS to perform well, how do we measure this, and how do we use this information to build a more successful system?

CURRENT EVALUATION MODELS

In the area of information retrieval (IR), evaluation has a long history that can be traced back to the automatic indexing studies pioneered by librarian and computer scientist Cyril Cleverdon at Cranfield University in the 1960s.¹ The basic IR evaluation model has been extended by efforts associated with the Text Retrieval Conference (TREC), an annual meeting cosponsored by the National Institute of Standards and Technology and the US Department of Defense that began in 1992.²

Basic IR model

In the basic IR evaluation model, researchers share test collections that contain a corpus, queries, and relevance assessments that indicate which documents are relevant to which queries. Because researchers share common resources and guidelines for conducting system evaluations, it is possible to compare search systems and improve search algorithms.

Particular evaluation measures indicate how well a search algorithm performs with respect to the number of relevant documents retrieved along with the position of these documents within a ranked list. Common measures

include precision, recall, mean average precision, mean reciprocal rank, and discounted cumulative gain.

While researchers explore different problems and search strategies, the basic objective of system evaluation is to assess search performance, which is usually tied directly to how effectively the system retrieves and ranks relevant information objects. This evaluation model abstracts away the human elements and focuses primarily on the topical relevance of documents to queries.

Interactive IR model

While most IR evaluations consider search performance, interactive IR focuses on how people use systems to retrieve information.³ IIR is informed by many fields including traditional IR, information and library science, psychology, and human-computer interaction (HCI). The TREC Interactive Track formalized the IIR evaluation method, and it has become the standard for laboratory evaluation of systems designed to support interactive IR.⁴

In a typical IIR evaluation, searchers use one or more experimental IR systems to find information described by a small number of prescribed topics—often using the IR evaluation test collections. Searchers' interactions with systems are logged, and at various points they provide feedback via questionnaires and other self-report techniques.

Typical outcome measures are usability and performance. While usability measures are based on searchers' responses to questionnaire items or their interactions with the system, performance measures are based on the number of relevant documents searchers find and the time it takes them to do so. Performance is often computed by comparing searchers' relevance assessments with baseline relevance assessments obtained from the test collection.


Other evaluation models

Both the IR and IIR models rely on laboratory evaluation, but researchers in these fields also conduct evaluations in real-world environments. Those working at search engine companies, for instance, can analyze search logs that contain billions of records. In some cases, researchers can conduct live trials of experimental algorithms or search interfaces by making them available to a subset of searchers.

While large-scale log studies let researchers observe numerous searchers with a diverse range of interests and information needs, these observations are limited to what can be captured in a search log, which primarily consists of queries and clicks. Important information about searchers, their information needs, and the basis for their actions is missing from such logs.

Ethnographic-style evaluations can lead to a more detailed understanding of searchers' real information needs,

but these are less common. Moreover, data can only be collected from a small number of searchers about a small number of tasks, which limits the generalizability of such studies. However, researchers can combine ethnographic observations with laboratory studies and log analyses to provide a richer picture of how systems support the search process.⁵



Important information about searchers, their information needs, and the basis for their actions is missing from large-scale log studies.

WHY CURRENT MODELS ARE INSUFFICIENT

Regardless of the location—laboratory or real world—current frameworks are insufficient for evaluating ISSSs for many reasons. First, the user and task models in traditional IR evaluation don't capture all types of information-seeking tasks, activities, and situations. Second, the Web's dynamic nature continually changes the base of objects available for retrieval over time. Third, information-seeking tasks are often complex and evolve without having stable, definable end points. Finally, information seeking occurs over sustained periods, implying the need for longitudinal evaluation designs that measure change.

Inadequate user and task models

Underlying all IR evaluations is some user model—an abstract representation of target searchers—and one or more task models, which represent user goals. The user and task models help define the particular behaviors and activities the system is intended to support and help determine the appropriateness of particular evaluation methods, measures, and searchers.

User models. These are often limited to experienced searchers with clearly defined search tasks. One common user model is that of a librarian or other search intermediary. Other examples include an intelligence analyst, patent searcher, or novice searcher. The general Web user can also function as a user model, although researchers often narrow the possibilities by including contextual characteristics such as how much a person knows about a topic or how much time is available to complete a task. However, there are few Web user models and many types of Web searchers, so more nuanced models are necessary for ISSS evaluation.

Task models. These include a wide range of search goals such as finding documents for a survey article, navigating to a key resource or homepage, checking a fact, and answering a question. The "Designing Exploratory Search Tasks for ISSS Evaluation" sidebar describes four criteria

➔ DESIGNING EXPLORATORY SEARCH TASKS FOR ISSS EVALUATION

Bill Kules, *The Catholic University of America*

Robert Capra, *University of North Carolina at Chapel Hill*

Exploratory search tasks arise from information needs in which users “lack the knowledge or contextual awareness to formulate queries or navigate complex information spaces, the search task requires browsing and exploration, or system indexing of available information is inadequate.”¹ They comprise an important class of information problems that share uncertainty, ambiguity, and discovery as common aspects.

Whether conducting usability studies or controlled experiments, researchers and practitioners must know that the tasks they use are ecologically valid and represent real-world information needs. Designing tasks to study exploratory search rather than a directed style of search can be especially difficult. At the same time, the tasks must be constructed in such a way that different research groups can compare the results between subjects in a single study and across multiple studies.

We determined that an exploratory search task

- indicates uncertainty, ambiguity in information need, or a need for discovery;
- suggests a knowledge acquisition, comparison, or discovery task;
- provides a low level of specificity about the information necessary and how to find the required information; and
- provides enough imaginative context for the study participants to be able to relate and apply the situation.

Based on these characteristics, we proposed a formal procedure for constructing tasks² that draws task topics from query log data, integrates them into a high-level work scenario,³ and addresses practical issues encountered in controlled or semi-controlled evaluations. An experimental evaluation of four tasks created using this procedure suggests that it led to well-grounded, realistic tasks that elicited exploratory search behavior.

for exploratory search tasks along with an example task. Most traditional IR evaluations are designed around a single search task that can be resolved in one session. It’s important to emphasize that while task models might specify how the desired information will be used, IR evaluation focuses on information finding. However, search is often not an end unto itself but a task conducted to achieve a larger goal.

Task models are also somewhat self-contained. It’s assumed that tasks are solvable in a single search session and, in IR evaluation, that a single query can adequately represent a searcher’s information need. Many information-seeking tasks require searchers to engage in multiple search sessions. During each session searchers may enter many queries and review several search results lists. However, most evaluation measures are based on a single query and a single list of ranked results and do not adequately capture searchers’ behaviors.

The following example task meets our exploratory search criteria and is based on a topic extracted from actual query logs from an online library catalog.

Imagine you are taking a class called “Feminism in the United States.” For this class you must write a research paper on some aspect of the US feminist movement, but have yet to decide on a topic. Use the catalog to find two possible topics for your paper. Then use the catalog to find three books for each topic so that you can make a decision regarding which topic to write about.

This task gives participants some direction—two topics, three books on each—but leaves the main aspects of the information need open for the participant to explore.

The annual Text Retrieval Conference (TREC) has demonstrated the value of well-constructed, comparable tasks in the evaluation of information retrieval systems. With growing interest in exploratory search from both researchers and practitioners, there is a need to develop such tasks that can be used in the study of exploratory search behaviors and systems.

References

1. R.W. White et al., “Supporting Exploratory Search,” *Comm. ACM*, Apr. 2006, pp. 36-39.
2. B. Kules and R. Capra, “Creating Exploratory Tasks for a Faceted Search Interface,” *Proc. 2nd Workshop Human-Computer Interaction and Information Retrieval (HCIR 08)*, Microsoft Research, 2008, pp. 18-21.
3. P. Borlund, “The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems,” *Information Research*, Apr. 2003; <http://informationr.net/ir/8-3/paper152.html>.

Bill Kules is an assistant professor in the School of Library and Information Science at The Catholic University of America. Contact him at kules@cua.edu.

Robert Capra is a research scientist in the School of Information and Library Science at the University of North Carolina at Chapel Hill. Contact him at rcapra3@unc.edu.

More recently, researchers have developed measures like session-based discounted cumulative gain to summarize performance in search tasks where multiple queries are used, results are of different quality and novelty, and stopping criteria vary.⁶ Such evaluation measures are important for characterizing information-seeking tasks because they more closely model searchers’ information-seeking behaviors.

People engage in information-seeking tasks for many reasons: to investigate curiosities, learn about some topic of interest, make connections between topics, stimulate creativity, and even for entertainment purposes. When such tasks are the goals, measuring task outcomes is difficult. How can we determine if someone has learned something by using an ISSS? How much learning is required to say the ISSS is effective? What does it mean for an ISSS to help satisfy a person’s curiosity or stimulate creativity? Appropriate outcome measures vary and are

tied directly to the task the user is trying to accomplish. Thus, the development of richer task models for ISSSs and corresponding evaluation measures are important research directions.

Dynamic test corpora

An issue related to user and task models is test corpora. A corpus is a set of documents, or information objects, that searchers access during a study. In traditional IR evaluations, test corpora are fixed and stable. Static corpora facilitate evaluation, as all systems are working with the same collection of information objects. Moreover, they make it possible to create topics that can be searched successfully and provide researchers with some information about the number of topically relevant documents.

Test corpora usually consist of mostly newswire text. Additional corpora that contain hyperlinked text and alternative types of information objects such as webpages, intranet pages, blog postings, or images have also been developed, but test corpora usually contain only one type of information object. This is quite different from typical information-seeking environments, where searchers are likely to encounter a variety of document types and genres of varying quality that constantly change over time.

Task complexity and evolution

There are many models of the information-seeking process, but for the most part they haven't made their way into IR system development and evaluation.^{7,8} These models characterize information seeking as a process that occurs across many search episodes using many different resources. Information seeking is interwoven with numerous other activities, and searchers commonly engage in multiple information-seeking tasks simultaneously.

These models also depict searchers as employing various information-seeking strategies, such as browsing related documents, that go beyond simply typing text into a query box and reviewing a list of search results. For instance, Marcia J. Bates' berrypicking model⁷ presents queries that change as the user engages in information seeking. This differs from the static and unitary query model assumed in traditional IR evaluation. Bates further posits that information needs are often not satisfied by a single, final retrieved set of documents but by a series of queries, navigations, and selections that occur throughout the information-seeking process.


Information-seeking tasks are often complex and evolve over time. While there might be objective, definable solutions for traditional IR search tasks, this isn't necessarily true for information-seeking tasks.

In addition, the information-seeking process itself is just as important—if not more so—than the final state. This suggests that traditional evaluation measures based

on system performance, more specifically on how many relevant documents are returned in response to a single query at a single point in time, must be extended. It also implies that the notion of topical relevance, which measures whether the document is topically related to the query, must be extended. Other types of relevance—situational, cognitive, and motivational—will become increasingly important in IR evaluation.⁹

The practice of evaluating performance using benchmark assessments based on objective, topical relevance is also no longer sufficient. Such assessments don't generalize across searchers, and it's difficult to create benchmark judgments based on other types of relevance because they're even more individualistic by nature.

Finally, the practice of asking searchers to make absolute relevance judgments of information objects becomes less useful since relevance assessments change throughout the course of information seeking.



Searchers commonly engage in multiple information-seeking tasks simultaneously.

Need for longitudinal designs

A limitation of the traditional IIR evaluation model is that it considers only a small slice of the search process—that which can be captured during a short experimental session. For some types of tasks this isn't too problematic. For example, many high-precision search tasks that require searchers to find answers to specific questions—say, the current weather, movie show times, or a celebrity's birth date—can be completed quickly.

The resolution of such tasks might be motivated by another larger task—for example, a user might search for show times because he plans to attend a movie—but these larger tasks usually don't require synthesizing and integrating information from multiple sources; rather, information needs are temporarily fixed and their resolution immediate. A system's ability to resolve such tasks is also clear to searchers, who can look at a results list and determine whether the information is there and, if so, where it is in the ranked list. However, the answers to such questions often change frequently, and their correctness may depend on when the searcher asks the question—for example, in looking for current weather conditions.

In the case of more open-ended information-seeking tasks, the temporal constraints of the traditional IIR evaluation model are more problematic as such tasks might not be resolved quickly and might be part of an ongoing quest for information that has no easily identifiable point of completion. This suggests the need for longitudinal study designs that let researchers observe series of information-seeking

➔ THE ISSS MEASUREMENT DILEMMA

Elaine G. Toms, *Dalhousie University*

Heather O'Brien, *University of British Columbia*

For more than half a century, precision, recall, and their variants were the standard norms for evaluating information retrieval systems. With the emergence of human-computer interaction, it became imperative to view IR systems through users' eyes to assess efficiency, effectiveness, and other perceptions. Researchers accordingly developed objective and subjective metrics that relate to users and the search context. However, such metrics—including query size, time on task, and satisfaction—tend to be measured independently despite the fact that IR systems have multiple interrelated components that control, manage, and affect user interactivity. The dilemma for information-seeking support system researchers is how to assess ISSSs given the complex nature of both the system and the human environment in which they operate.

The first metric to assess relationships among aspects of a complex IR system was "informativeness," which combined a subjective user response regarding usefulness of the information retrieved with the system's ability to present relevant items in the most useful (to the user) order.¹ Based on sound mathematical principles and information search theory, this metric was the first to measure user interactivity in IR systems.

The limited availability of complex metrics dictates the need for methods that examine interrelationships among multiple simple metrics. Two such techniques that are not commonly used to evaluate IR systems are factor analysis and structural equation modeling. Both FA and SEM enable researchers to examine different types of data—user attitudes, observed behaviors, system performance, and so on—simultaneously for a more holistic approach. Because IR systems are tools, their success is tied intrinsically to human use. A system cannot be assessed independently of its user; this calls for integrated metrics that reflect interactivity.

FA looks for simple patterns in the associations among variables and extracts the most parsimonious set. For example, we used FA to map metrics to dimensions of relevance, deducing that three core factors or variables—system, user, and task—could be measured with eight objective or subjective metrics of user and system performance.²

SEM goes a step further, combining confirmatory FA with path analysis to confirm factors and build predictive models about

their relationships. Because SEM models are theoretically derived, data analysis tests the "fit" between the hypothesized and actual relationships in the data, indicating which variables are independent and which are directly or indirectly related to a larger set. We used SEM to evaluate a six-factor scale of user engagement,³ confirming both the presence of the factors—aesthetics, novelty, involvement, focused attention, perceived usability, and durability—and the predictive relationships among them.

ISSSs are complex systems with multiple features that enable multiple types of interactivity. Techniques such as FA and SEM facilitate the assessment of varied, multiple, simple measures. Core to these approaches is the need for a clear theoretical focus on measurement selection and interpretation of the output, much like any other statistical technique. Both will fail if the user "dumps" in a set of data and makes sweeping conclusions about the results, independent of the theoretical foundation that informs the phenomenon under examination. Used appropriately, however, FA and SEM could lead to the creation of complex metrics for a more holistic evaluation of ISSSs.

References

1. J. Tague-Sutcliffe, "Measuring the Informativeness of a Retrieval Process," *Proc. 15th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 92)*, ACM Press, 1992, pp. 23-36.
2. E.G. Toms et al., "Searching for Relevance in the Relevance of Search," *Information Context: Nature, Impact, and Role*, F. Crestani and I. Ruthven, eds., LNCS 3507, Springer, 2005, pp. 59-78.
3. H. O'Brien, "Defining and Measuring Engagement in User Experiences with Technology," doctoral dissertation, Dalhousie University, 2008.

Elaine G. Toms is Canada Research Chair in Management Informatics and an associate professor in the Faculty of Management at Dalhousie University, Halifax, Nova Scotia, Canada. Contact her at etoms@dal.ca.

Heather O'Brien is an assistant professor in the School of Library, Archives, and Information Studies at the University of British Columbia, Vancouver, British Columbia, Canada. Contact her at hlobrien@interchange.ubc.ca.

activities that occur over time. These types of study designs are more time-consuming and require researchers to give up some experimental control.

In laboratory IR evaluations, many of the variables that are not of immediate interest are controlled. In ISSS evaluation, however, searching occurs in a much richer context, making control more difficult and less desirable in many cases. While traditional evaluation has focused on component analysis, which allows isolation and control of variables, holistic evaluation models are needed to capture more of the variability in ISSSs. Longitudinal evaluation models require more sustained engagement with searchers, the development of a wider range of less intrusive instruments for data collection, and richer analysis

methods for identifying important information-seeking behaviors and outcomes. The sidebar "The ISSS Measurement Dilemma" describes in more detail the need for combining complex metrics.

Web search log studies are another example of longitudinal, sustained engagement and use of unobtrusive data-collection techniques. However, many of these logs contain only partial information about user behavior. Those that rely on server-side logging capture the user's communication with one particular search service but not what searchers do after they navigate away from the service. Client-side logging captures a more complete picture of searchers' behaviors and can be instrumented via Web browser toolbars, but researchers have the added chal-

lenge of identifying information-seeking behaviors amidst the many other recorded actions. Regardless of what type of logging is used, there's a need to collect additional data that provides a context for interpreting searchers' actions in these logs.

EVALUATION DIRECTIONS

Information system development and evaluation are complex and usually require interdisciplinary teams including engineers, computer scientists, and behavioral scientists. The analysis of information seeking and exploration, and the development of systems to support these activities, can be informed by many different scholars with a wide range of expertise. Thus, creating a diverse research community is necessary for significant advancement.

While the research and development of information-seeking models have paralleled IR research, these two areas have largely emerged independently. IR evaluation has been driven by Cranfield-style retrieval experiments and user studies most often conducted in laboratory environments, which require a certain amount of abstraction and control. In contrast, information-seeking models have primarily emerged through naturalistic, qualitative studies involving small sets of searchers.

ISSS represent an opportunity to integrate these models and create a new framework for evaluation. They can also broaden and extend information-seeking models through large-scale evaluation. For instance, these models have typically excluded representation of the information environment; future models must accommodate the variety of environments in which information seeking occurs.

One way to create a research community that includes participation from the broad range of disciplines needed to develop and evaluate information-seeking systems is to create shareable resources that facilitate and enable participation. Such resources should contain datasets; search components; data-collection tools, methods, and measures; and operational search environments in which to explore new ideas.

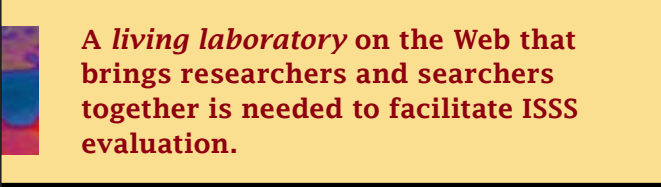
Datasets

Potentially valuable datasets include both large-scale Web log data and data from more controlled laboratory studies. The former contain millions of search records contributed by thousands of users. These datasets capture wide, geographically distributed segments of the population, and the sheer volume of data allows for greater generalizability and the development of powerful models. Those from smaller, focused studies often have richer contextual data, and much more information is available about what searchers are trying to accomplish. Interviews and video recordings of the screen often accompany these types of datasets. A repository of datasets—with appropriate consent from and privacy protection for

searchers—would provide a point of collaboration for researchers and allow examining and analyzing data using a wider array of methods and techniques.

Search components

Search components that can easily be plugged into experimental systems or used to examine behavioral issues must also be developed. ISSSs will likely perform various functions that extend beyond search, although this will remain an important component. However, it's very difficult for individual researchers, or even small teams of researchers, to develop an end-to-end system that is stable, robust, effective, and efficient. Moreover, because components work together, one that doesn't work adequately will impact searchers' experiences with the entire system.



A living laboratory on the Web that brings researchers and searchers together is needed to facilitate ISSS evaluation.

Achieving a stable working system for IIR research requires substantial engineering effort. This presents a high barrier to entry to researchers—especially those in traditional social science disciplines, who may not have expertise in search technology or in building and deploying large-scale systems, but may have considerable knowledge about human behavior and information processing—and greatly limits the number of iterations that can be performed. Developing resources that support richer collaborations and contributions from the wide range of relevant disciplines is a key enabler for improving information-seeking support and for developing a richer theoretical basis for ISSS work.

Tools, methods, and measures

Other needed types of shared resources are those that let researchers collect data from searchers. This includes loggers that monitor user interactions—with Web search engines, vertical search engines, browsers, and so on—as well as capture page contents. In addition to traditional loggers, data-collection instruments are needed that enrich log data. These might elicit information from searchers about their goals, needs, and states at various points in time. Such data could supplement log data and provide more information-seeking context.

New evaluation methods and measures are also top priorities for ISSS evaluation. Searchers seeking to sustain an interest may have different expectations of how the system should support them than searchers with more clearly defined goals such as finding relevant documents or specific answers to questions. Understanding these


expectations will likely lead to additional evaluation criteria that reflect success. In addition, more process-specific measures are needed that capture learning, cognitive transformation, confidence, engagement, and effect. Performance will still be important, of course, but measures are needed that don't depend on benchmark relevance assessments and that consider multiple query iterations and search sessions.

Studies that seek to describe the range of information-seeking tasks, processes, and strategies in which searchers engage are also necessary. These studies can help establish user and task models for more focused evaluations, such as those that might occur in a laboratory, and help developers understand the behaviors and activities that ISSSs must accommodate. Conceptualizing tasks and creating task models, in particular, are very important activities since they determine appropriate evaluation measures.

Evaluation environments

Researchers traditionally work relatively independently building infrastructure and tools. For each study, they consult various sources to gather collections and search tasks. Recruiting searchers also presents challenges, and researchers are often limited to a particular type of searcher that is nearby and easy to contact.

A *living laboratory* on the Web that brings researchers and searchers together is needed to facilitate ISSS evaluation. Such a lab might contain resources and tools for evaluation as well as infrastructure for collaborative studies. It might also function as a point of contact with those interested in participating in ISSS studies. For instance, many IR researchers use crowdsourcing via the Amazon Mechanical Turk as a way to obtain relevance assessments.¹⁰ Such techniques can also be used to solicit research participants for ISSS evaluation. The development of an infrastructure for facilitating recruitment, retention, and participation will let researchers expand the range of participants in their studies and ultimately broaden knowledge about ISSSs.

ISSSs provide an exciting means to extend traditional IR and IIR evaluation models, and to create a research community that embraces diverse methods and participation. An opportunity also exists for incorporating more aspects of the information-seeking process into ISSS development and evaluation, and for building a richer theoretical foundation. Community participation and shared resources are keys to leveraging existing expertise, as well as attracting additional participants who can broaden perspectives and enhance understanding of the information-seeking process and the systems needed to support it. 

References

1. C. Cleverdon, "The Cranfield Tests on Index Language Devices," *Readings in Information Retrieval*, K.S. Jones and P. Willett, eds., Morgan Kaufmann, 1997, pp. 47-59.
2. E.M. Voorhees and D.K. Harman, eds., *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, 2005.
3. I. Ruthven, "Interactive Information Retrieval," *Ann. Rev. Information Science and Technology*, vol. 42, 2008, pp. 43-91.
4. S.T. Dumais and N.J. Belkin, "The TREC Interactive Tracks: Putting the User into Search," *TREC: Experiment and Evaluation in Information Retrieval*, E.M. Voorhees and D.K. Harman, eds., MIT Press, pp. 123-153.
5. C. Grimes, D. Tang, and D.M. Russell, "Query Logs Alone Are Not Enough," *Proc. Workshop on Query Log Analysis at the 16th Int'l World Wide Web Conf.*, ACM Press, 2007; www2007.org/workshops/paper_51.pdf.
6. K. Järvelin et al., "Discounted Cumulated Gain-Based Evaluation of Multiple-Query IR Sessions," *Advances in Information Retrieval*, LNCS 4956, Springer, 2008, pp. 4-15.
7. M.J. Bates, "The Design of Browsing and Berrypicking Techniques for the Online Search Interface," *Online Rev.*, Oct. 1989, pp. 407-424.
8. C.C. Kuhlthau, *Seeking Meaning: A Process Approach to Library and Information Services*, 2nd ed., Libraries Unlimited, 2003.
9. T. Saracevic, "Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance," *J. Am. Soc. Information Science and Technology*, Nov. 2007, pp. 1915-1933.
10. O. Alonso, D.E. Rose, and B. Stewart, "Crowdsourcing for Relevance Evaluation," *ACM SIGIR Forum*, Dec. 2008, pp. 9-15.

Diane Kelly is an assistant professor in the School of Information and Library Science at the University of North Carolina at Chapel Hill. Her research interests are in interactive information retrieval and user behavior, and evaluation methods and metrics. Kelly received a PhD in information science from Rutgers University. Contact her at dianek@email.unc.edu.

Susan Dumais is a principal researcher and manager of the Context, Learning, and User Experience for Search group at Microsoft Research. Her research interests are in information-retrieval algorithms, interfaces, and evaluation. Dumais received a PhD in cognitive psychology from Indiana University. She is an ACM Fellow and a member of the CHI Academy. Contact her at sdumais@microsoft.com.

Jan O. Pedersen is the chief scientist at A9.com, an Amazon company. His research interests span the range of Web search technologies, including matching, ranking, and evaluation. Pedersen received a PhD in statistics from Stanford University. He is an ACM Distinguished Scientist. Contact him at jpederse@yahoo.com.