Report on ACM SIGIR 2006 Workshop on Evaluating Exploratory Search Systems

Ryen W. White[1], Gheorghe Muresan[2], and Gary Marchionini[3]
[1] Microsoft Research, One Microsoft Way, Redmond, WA 98052 USA
[2] School of Communication, Information and Library Studies,
Rutgers University, New Brunswick, NJ 08901 USA
[3] School of Information and Library Science,
University of North Carolina, Chapel Hill, NC 27599 USA
ryenw@microsoft.com, muresan@scils.rutgers.edu, march@ils.unc.edu
Workshop website: http://research.microsoft.com/~ryenw/eess

Abstract

Exploratory search systems (ESS) are designed to help users move beyond simply finding information toward using that information to support learning, analysis, and decision-making. The evaluation of the interactive systems designed specifically to help exploratory searchers is a challenging area, worthy of further discussion in the research community. In this article we report on a workshop conducted in conjunction with the ACM SIGIR Conference in Seattle, USA, in August 2006. The workshop involved researchers, academics, and practitioners discussing the formative and summative evaluation of ESS.

1 Background and Introduction

People engaged in an exploratory search (ES) require support that goes beyond the known-item retrieval well-handled by many modern search systems. They may need help discovering new associations and kinds of knowledge, resolving complex information problems, or developing an understanding of the terminology and the information space structure [9]. Whilst search is expanding beyond supporting simple lookup into supporting complex behaviors involving information-seeking and information use, there is no framework for evaluating this new genre of search system. The annual NIST-sponsored Text Retrieval Conference (TREC) has provided a medium for the evaluation of algorithms underlying the analytic aspects of Information Retrieval (IR) systems, yet struggled because the experimental methods of batch retrieval are not suited to studies of interactive search. Since TREC-3, the conference has extended its mandate to recognize the role of the user IR activities. The Interactive Track [5], and later the HARD track [1], have both attempted to bring the user into the loop. However, these tracks struggled to establish comparability between experimental sites, in terms of the experimental systems devised, the research hypotheses tested and the measures used. They were also adversely affected by the dependence on relevance judgments and interactions between users, tasks, and systems. Nonetheless, the Interactive Track was successful at highlighting the importance of users in IR research [7].

The creation of *exploratory search systems* (ESS) to support exploratory search activities has become a topic of great interest to researchers [9]. The research community has focused for some time on how to

develop novel interfaces to support users engaged in exploratory search. However, given the range of ESS now available, it seems time to shift the focus of research toward understanding the behaviors and preferences of searchers engaged in exploratory searching, on tasks supported by such systems, and on measuring exploration success. For example, a key component of exploration is human learning (a topic studied extensively by cognitive psychologists [8]) yet this issue has not been explored in relation to ESS. Our vision is of a framework for ESS evaluation that could validate the support these systems offer, and chart new courses toward improved search experiences for users. The workshop we describe in this report is the first step in that direction. It brought together researchers from communities such as information retrieval, library and information science, and human-computer interaction, for a discussion of the issues related to the formative and summative evaluation of ESS.

2 The Workshop

Around 35 researchers with a broad range of experience and expertise attended the workshop. The day began with a short ice-breaking activity to familiarize participants with the goals of the workshop and with each other. A four-person panel on ESS evaluation issues, chaired by an organizer, served as an introduction to discussing exploratory search, and the challenges of evaluating exploratory search systems. The panel was followed by paper presentations, breakout sessions, and related discussions. We will now provide more details on the main activities of the day.

2.1 Introductory Activity

Following a brief welcome by the organizers, Ryen White led the participants in an introductory activity that was devised to familiarize them with the topic and with each other. Participants were divided into two groups based on their seating locations and each group was assigned a question to discuss:

- Group 1: “Why is evaluating exploratory search systems challenging?”
- Group 2: “How can we evaluate these systems?”

Prior to the group assignment, the organizers attached three headings to each of two walls in the classroom in which the workshop was being conducted. These headings were:

- Methodologies (e.g., test collections, simulations, ethnography, task-oriented approaches)
- Metrics (e.g., measures of learning, user performance)
- Models (e.g., interfaces and interaction paradigms, mental models)

Participants in each group were given sticky notes and asked to discuss their assigned question amongst themselves. They were instructed to place notes containing ideas that emerged during their discussion under the appropriate heading on the wall. This was felt to be a useful ice-breaking exercise to get participants to interact and share ideas early in the event. As an additional benefit, the notes that emerged from this activity were used in the break-out sessions later in the day.

2.2 Panel

A four person panel – chaired by Gary Marchionini – followed the introductory activity. One of the aims of this panel was to get some insights from experts that attendees might not know or might not have a chance to hear from regularly. We hoped that doing this would broaden the discussion and make for a more rewarding event. Since exploratory search is closely related to research in the Human-Computer Interaction and Psychology communities, we invited panelists better known for their research in those areas, with experience in evaluating interactive systems in a variety of ways. The panelists and the titles of there presentations were:

- **Andy Edmonds**, Windows Live Search
  Building Models of Search Success with Experience Sampling and Event Logs
- **Cathy Marshall**, Microsoft
  Why a Corpus-Topics-Relevance Judgments Framework Isn't Enough: Two Simple Retrieval Challenges From the Field
- **Thomas Landauer**, Pearson Knowledge Technologies / University of Colorado
  Retrieval Evaluation sans Human Relevance Judgments
- **Peter Pirolli**, Xerox PARC
  Analysis of the Task Environment of Sense Making

All panelists gave presentations of relevant aspects of their research, answered questions from attendees, and participated in discussions, and other workshop activities throughout the day. Edmonds discussed the use of an experience sampling methodology [2] that models user feedback at critical points during the search (e.g., leaving a search result page, returning to result page via the back button), in order to generate predictions of searcher satisfaction in other uninterrupted search sessions. He described research involving the combination of explicit feedback (elicited via dialog boxes) and implicit feedback (mined from interaction logs) to create predictive models of search. Edmonds concluded by suggesting that factors such as engagement, learning and user workload, quality of the results selected, and the identification of user intent (implicitly and explicitly) should be considered when evaluating exploratory search systems.

Marshall discussed her research in context and personal memory, with particular focus on retrieval tasks falling outside the traditional corpus-topics-relevance judgment evaluation framework traditionally used in IR evaluation. She focused on two challenges in particular: (i) encountering new information or re-encountering forgotten information, and (ii) the retrieval of an appropriate version of semi-redundant information, either from a personal information space or from a public store of datasets. Marshall suggested that there was a need to examine the assumptions that underlie TREC and similar evaluation paradigms, because they do not consider many real-world activities attributable to human failings or information sloppiness, those that go beyond the information retrieval rubric (which generally focuses on the presence of an information need), and those that are part of invisible information interaction processes (such as the creation of experimental datasets).

Landauer described his experience in using Latent Semantic Analysis (LSA) to extract gist from essays, student papers, and formal publications. He described a novel automatic retrieval evaluation method for LSA-based cross-language retrieval that does not require the creation of example queries or human relevance judgments, and produces a quantitative distribution of expected performance over all possible queries within a corpus. He suggested that the technique can provide rapid feedback on the design on IR systems under development, or as a means of testing operational systems.

Pirolli discussed how cognitive analyses of human behavior can be leveraged to assess exploratory search, with particular emphasis on models of expertise. He proposed an initial sketch of a measurement framework and theory for evaluating exploratory search centered on the notion of task environments (i.e., the physical, social, virtual, and cognitive environments that drive human behavior), and the use of cognitive task analysis to shape the framework and suggest points of measurement within it. As an example, Pirolli presented a study of experts in intelligence analysis that revealed the presence of *dual space search* [6] whereby users first explore, enrich and exploit (foraging), then participate in activities involving problem structuring, evidentiary reasoning, and decision making (sensemaking).

The panel gave us a unique set of perspectives on the challenges of ESS evaluation, and we are grateful to them for taking the time to attend.

2.3 Paper presentations

Those interested in participating in the workshop were required to submit a short position paper describing their research. Papers were solicited in areas ranging from human learning and mental models for exploratory search processes, to the role of context, test collections, and ethnography and field studies. Submitted papers were reviewed by an international program committee who were asked to make a recommendation for acceptance/rejection and provide comments that authors could use when creating the final version of their papers for the working notes distributed to attendees. The working notes are available on the workshop website: http://research.microsoft.com/~ryenw/eess.

There were a total of 18 papers submitted to the workshop. We accepted 11 papers: five as background to appear in the working notes, and six papers for oral presentation *and* to appear in the working notes. The titles and authors of the accepted papers are shown in Table 1.

Table 1. List of accepted papers.

Accepted for oral presentation	Accepted as background
• ***Layered Evaluation of Adaptive Search.*** P. Brusilovsky, R. Farzan, J.-W. Ahn (U. Pittsburgh, USA) • ***Wrapper: An Application for Evaluating Exploratory Searching Outside of the* Lab.** B. J. Jansen, R. Ramadoss, M. Zhang, N. Zang (Penn. State U., USA) • ***Exploratory Search Visualization: Identifying Factors Affecting Evaluation.*** S. Koshman (U. Pittsburgh, USA) • ***Task-based Evaluation of Exploratory Search Systems.*** W. Kraaij, W. Post (TNO, The Netherlands) • ***Methods for Evaluating Changes in Search Tactics Induced by Exploratory Search Systems.*** B. Kules (Takoma Software, USA) • ***An Integrated Approach to Interaction Modeling and Analysis for Exploratory Information Retrieval.*** G. Muresan (Rutgers U., USA)	• ***Exploratory Search in Wikipedia.*** S. Fissaha, M. de Rijke (U. Amsterdam, The Netherlands) • ***A Pilot for Evaluating Exploratory Question Answering.*** V. Jijkoun, M. de Rijke (U. Amsterdam, The Netherlands) • ***Impact of Relevance Intensity in Test Topics on IR Performance on Polyrepresentative Exploratory Search Systems.*** B.R. Lund, P. Ingwersen (Royal School of Library and Information Sciences, Denmark) • ***From Question Answering to Visual Exploration.*** D. McColgin, M. Gregory, E. Hetzler, A. Turner (Pacific Northwest National Laboratory, USA) • ***What do the Attributes of Exploratory Search Tell us about Evaluation.*** Y. Qu, G.W. Furnas (U. Michigan, USA)

The six papers accepted for presentation focused mostly on tools and methodologies for evaluating exploratory searching. Brusilovsky and co-authors argued in favor of a layered approach to evaluation, in order to better estimate the quality of the different components of a retrieval system. They separated the decision-making layer, which analyzes the interaction history of the user's social group and predicts document relevance, from the adaptation layer, which uses visual cues to convey information about the search hits displayed. While the reported results indicate the effectiveness of an adaptive retrieval system based on social search, the more important conclusion is that the layered approach to evaluation is effective, as is being able to pinpoint the effect of the system's different components.

When designing and developing the *Wrapper* system, Jansen *et al.* attempted to address realistic information seeking scenarios, in which multiple software tools may be used for planning searches, possibly in multiple search episodes, for finding relevant documents, extracting information, writing notes or annotations, etc. Based on a client-server architecture, their tool logs a wide variety of user-browser interactions and monitors the user's access to a variety of information sources. The reported user study demonstrated that, in a naturalistic setting, users engage in a wide variety of information-

seeking activities from various sources, which would not be captured by traditional, system or server based logging.

Koshman summarized results from a number of user studies in which the search interfaces were based on information visualization tools that supported the exploration of the information space, and of the relationships between search results, between search results and the search query, or between search results and the information sources. She suggested that, while visualization tools appear to enhance the functionality of systems and offer intuitive solutions for exploratory search, evaluation of such systems is made more complicated by the visualization tools: the effect of the cognitive load and the learning curve imposed by new mental models are confounded with the effect of the actual search algorithms or adaptation models. Based on the experience gained from these studies, Koshman made a number of recommendations for setting up future user experiments and suggested a number of variables that should be measured and considered when analyzing exploratory search.

Kraaij reviewed a number of interactive IR experiments well documented in the literature and argued that the adequacy of a system should be measured by its support for task completion and by user satisfaction, and that proper interaction design is more conducive to success than good indexing or search algorithms. While more time-consuming and more difficult to evaluate than "TREC-style" evaluation, user-centered task-based approaches are necessary in order to assess the real effectiveness of retrieval systems. Kules continued the same argument, emphasizing the need to combine lab experiments in controlled setting with naturalistic, longitudinal studies. While the former have the advantage of control, comparability and repeatability, the latter are essential in observing real search behaviors, and assessing the satisfaction or frustration of real people looking for information.

To conclude the presentations, Muresan described a methodology that integrates the design of the search interface, of the logging software and of the log analysis software, and relies on existing technology and open-source software. The essential step is the state-based design of the interaction, using the Unified Modeling Language (UML): the states of the user interface are identified, together with the user actions valid in each state and the state transitions triggered by these actions. While it is common to use states as a first step in designing the interface, the UML state diagram also supports deriving the XML schema of the interaction logs by specifying the actions and state transitions to be logged. Consequently, the code for logging events and the code for parsing the logs in view of analysis can be generated automatically, which improves development speed and accuracy. This methodology supports a wide variety of research questions, from quantitative analysis of logs to modeling search behaviors.

The background papers provide additional insight into current research in exploratory search. Fissaha and de Rijke investigated searches on Wikipedia and developed a questionnaire for eliciting search, discovery and retrieval requirements from Wikipedia users. Jijkoun and de Rijke developed a pilot for supporting answering of complex questions in Wikipedia in both monolingual and multilingual scenarios. Lund *et al.* looked at the effect that the density of relevant documents has on search effectiveness, and at its correlation with high overlap between multiple search engines. McColgin *et al.* developed an integrated question answering system that combines visual analytical tools with query expansion functionality, and is targeted at information analysts. Appropriately, their evaluation looked not only at the usability and utility of the interface, but also at its support for a variety of search strategies. Finally, Qu and Furnas discussed the interplay between searching, browsing, and the task representation. Based on the belief that the purpose of evaluation is to guide design, they paid particular attention to the tight coupling between search and other information activities, and at the gradual evolution of the task structure representation.

2.4 Breakout Sessions and Related Discussion

Following the paper presentations, participants were asked to form one group for each of the three subject headings used in the introductory activity. During the day, one of the organizers arranged the notes that had been attached to the wall into piles for the three headings: Models, Metrics, and Methodologies. Groups were asked to discuss the importance of each of these issues, and asked to elect a spokesperson to report back to the rest of the workshop. These reports served as a springboard for further discussion involving all attendees.

3 Discussion

The workshop succeeded in outlining the challenges of evaluating exploratory search systems and went some way toward addressing them. In this section we cover the main issues raised by the three breakout session groups:

3.1 Models

Models provide a foundation for thinking about problems and processes. The models breakout group first agreed that models for exploratory search are necessarily complex and dynamic, recognizing that disparate models lead to disparate evaluation paradigms, which in turn makes results across studies and settings difficult to compare. A variety of objects and processes were offered as fodder for discussion, including: human information processing (cognitive and affective) models, task models, communication models, models of human interests, models of learning, models of expertise, and models of search processes. Several participants argued that it is difficult to model exploratory search without grounding the model in content domains. As the discussion progressed, participants focused on learning in that exploratory search is often directed to learning tasks and the cognitive processes we observe as people explore information spaces reflect the same kind of behaviors observed in learning.

The group began to converge on multi-faceted models that combined several of the objects and processes noted on the sticky notes. Perhaps the most complete model integrated the information seeker, search task, information base, search system, and effects. Whether the information seeker or the task sits at the center of this model was not determined, but these elements seem to cover the main factors in exploratory search. In hindsight, we might consider adding the search context (including motivation, time and environmental constraints, etc.). Figure 1 represents the elements of this model with arrows representing the interrelationships. Double-sided arrows suggest two-way interactions as exploratory search progresses. The multiple arrows between the information base and system are meant to suggest that there are many possible paths that the information seeker may choose. The arrows from outcomes (effects or search results) back to task and information seeker suggest that there are multiple iterations over the exploratory session.

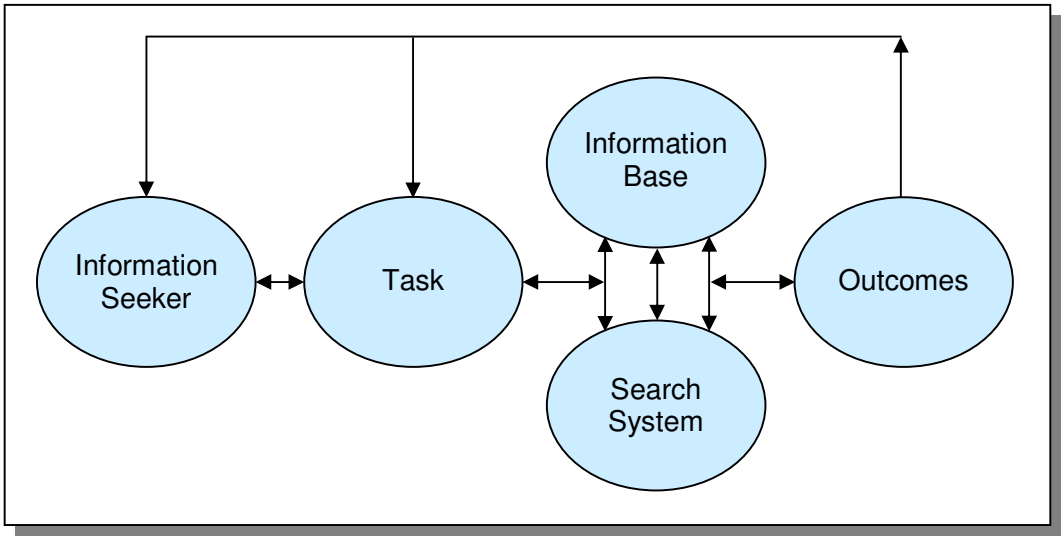

Figure 1. Model of exploratory search process.

3.2 Metrics

Evaluation metrics facilitate the incremental improvement of search technologies by providing a way to assess system performance and facilitate comparisons between experimental systems. Devising such metrics in exploratory search is particularly challenging since the goals are subject to change as the searcher interacts with the system. In such cases, standard precision-recall metrics may be ineffective (although there has been some consideration given to devising variants of these metrics more suitable for interactive information retrieval [3]). The metrics breakout group agreed that it is generally difficult to objectively assess the success of exploratory search sessions. The participants conducted a brainstorming session to identify a set of metrics (albeit at a fairly high level of abstraction) that could be used to test ESS, and proceeded to discuss each of these metrics.

Engagement and enjoyment: The degree to which users are engaged and are experiencing positive emotions can be a strong indicator of system performance. Given the amount of interaction required during exploration, the extent to which they are focused on the task and seem happy with the system's response can indicate whether it is fulfilling its role in supporting search activity. The number of actionable events (purchases, forms filled, bookmarking or feedback events, forwarding events etc.) could be used as a metric to approximate levels of engagement and enjoyment.

Information novelty: Since the goal of exploration is to encounter things not seen before, it seems appropriate to include the amount of new information encountered as a way of measuring the effectiveness of an ESS. Participants also saw the rate at which users encounter new information as an important aspect of novelty that may provide additional information on interaction throughput.

Task success: Group members agreed that task success should not only be based on whether the user reaches a particular target document, but also on whether they were able to encounter a sufficient amount of information of sufficient detail en route to reaching their goal. As one participant remarked "[exploratory search] is more about the journey than the destination". Since task success may be based on the difficulty of the task, metrics such as the Clarity measure [4] may also be appropriate.

Task time: The time spent to reach a state of task completeness was suggested as a good way to assess the efficiency of exploration activities. Suggestions included the total time spent, the time spent looking at irrelevant documents, and the proportion of time spent engaged in directed search versus the amount of time spent exploring. Participants suggested that task completeness would be indicated by experimental subjects based on their own perceptions of their task state.

Learning and cognition: Learning is key part of exploratory search. Through being able to measure cognitive and mental loads, the attainment of learning outcomes, the richness/completeness of a user's post-exploration perspective, the amount of the topic space covered, and the number of insights they come up with, group members suggested that we could compare ESS in terms of learning and cognition. To this end we could draw on much of the literature in the Human Computer Interaction, Cognitive Psychology, and Education communities, who have been addressing the measurement of similar metrics for some time.

3.3 Methodologies

The role of evaluation is primarily to assess the success of the information seeking process and, consequently, to inform the design of future retrieval systems. Therefore, evaluation methodologies are tightly connected to both the models chosen to represent the interaction and the metrics adopted for measuring success. Indeed, the models specify what factors of the interactions are most representative or relevant in a certain context, and metrics represent both a conceptualization of the models and a measure of retrieval success. As such, evaluation methodologies connect models and metrics by specifying the rules, methods and assumptions employed in evaluation, as well as the rationale and the philosophy behind the evaluation.

It is quite natural for an emerging research community or a new area of research to be tentative in developing or adopting an evaluation methodology, and that is the case with research in ESS. So far researchers have adopted experimental settings, document collections and investigation methods from areas such as Interactive Information Retrieval, Human Computer Interaction or Cognitive Psychology. The goal of reaching methodological rigor in studying exploratory search has not been reached but, as the field matures, a core of methods accepted by the research community is expected to crystallize. In order to reach these goals, the breakout group has agreed that a wide variety of candidate methods has to be employed, with the expectation that the best methods and practices will eventually prevail.

For example, there was agreement that naturalistic, longitudinal studies should be employed alongside lab experiments in controlled conditions. The former are better suited to observe the information seeker's information behavior and search strategies, as well as changes in information needs and behavior that occur in time. Moreover, they are invaluable in developing and testing interaction models, and making sure that assumptions used in user models do hold in general or in certain contexts. On the other hand, laboratory studies have the advantage of comparability and repeatability, and support quantitative studies that attempt to answer research questions about the level of support that different system components offer to the user.

Nevertheless, building appropriate test collections remain a difficult challenge: most existing experimental settings are based on the assigned task paradigm, and on the assumption that the information need is static during the interaction. More realistic settings should take into account the learning that takes place during the search session, the evolution of the information need, the dynamic nature of relevance judgments, as well as the personality, background, knowledge and preferences of the searcher.

3.4 Summary

To facilitate discussion in the workshop we divided participants into groups to talk about models, metrics, and methodologies separately. However, what emerged during the discussions was the importance of the interplay between these three aspects of evaluation. The success of the workshop has motivated us to continue our work in this area. We are currently organizing a special topic issue of *Information Processing and Management* on Evaluating Exploratory Search Systems, which will serve as an opportunity for authors from within the workshop and beyond to provide a more detailed description of their research in this area. If the submissions to the special issue are of similar caliber to the workshop submissions, presentations, and discussions, it should make for a worthwhile contribution in the rapidly emerging area of exploratory search. These activities are aimed at creating a community of interest in this area, and represent the initial steps toward the creation of a framework for the evaluation of ESS.

4 References

[1] Allan, J. (2003). HARD Track Overview in TREC 2003: High accuracy retrieval from documents. In *Proceedings of the Text Retrieval Conference*, pp 24-37.

[2] Barrett, L. F. and Barrett, D. J. (2001). An introduction to computerized experience sampling in psychology. *Social Science Computing Review*, 19(2):175-185.

[3] Borlund, P. and Ingwersen, P. (1998). Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 324-331.

[4] Cronen-Townsend, S., Zhou, Y., and Croft. W. B. (2002). Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 299–306.

[5] Dumais, S. and Belkin, N.J. (2005). The TREC Interactive Track: Putting the user into search. In Voorhees, E. and Harman, D. (Eds.) *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press.

[6] Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes.* Bradford Books, Cambridge, MA.

[7] Lagergren, E. and Over, P. (2001). Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 164-172.

[8] Landauer, T.K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *The psychology of learning and motivation*, 41: 43-84.

[9] White, R.W., Drucker, S., Kules, B. and schraefel, m.c. (2006). Supporting exploratory search. *Communications of the ACM (Special Section)*, 49 (4): 36-39.