Evaluating Exploratory Search Systems

Introduction to Special Topic Issue of Information Processing and Management

Ryen W. White
Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
ryenw@microsoft.com

Gary Marchionini
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
march@ils.unc.edu

Gheorghe Muresan
Microsoft Live Search
One Microsoft Way
Redmond, WA 98052 USA
ghemur@microsoft.com

Exploratory search can be used to describe an information-seeking problem context that is open-ended, persistent, and multi-faceted; and to describe information-seeking processes that are opportunistic, iterative, and multi-tactical. In the first sense, exploratory search is commonly used in scientific discovery, learning, and decision making contexts. In the second sense, exploratory tactics are used in all manner of information seeking and reflect seeker preferences and experience as much as the goal (Marchionini, 2006). In exploratory search people usually submit a tentative query to get them near relevant documents then explore the environment to better understand how to exploit it, selectively seeking and passively obtaining cues about where their next steps lie. Exploratory search can be considered a specialization of information exploration, a broader class of activities where new information is sought in a defined conceptual area; exploratory data analysis is another example of an information exploration activity. Exploratory search *systems* (ESSs) capitalize on new technological capabilities and interface paradigms that facilitate an increased level of interaction with search systems. Examples of ESSs include information visualization systems, document clustering and browsing systems, and intelligent content summarization systems. ESSs go beyond returning a single document or answer in response to a query, and instead aim to instigate significant cognitive change through learning and improved understanding. ESSs support aspects of *sensemaking* (Dervin, 1998) (i.e., through information visualization and other depictions they help create situational awareness and understanding in support of decision-making), and *information foraging* (Pirolli & Card, 1995), (i.e., they support the exploration and identification of information patches, and maximal information gain). For example, browsing is a serendipitous activity that can be attractive to users, who may benefit from the extraneous information (Marchionini & Shneiderman, 1988). ESSs help users engaged in browsing maximize their rate of information gain, make decisions about which navigational paths to follow, and understand the information they encounter. In addition, through interface features such as dynamic queries (Shneiderman & Plaisant, 2004) ESSs can help users see the immediate impact of their decisions.

In recent years researchers have focused on the development of new systems and interfaces to support exploratory search activities, not on their evaluation. It is necessary to shift the focus of research towards understanding the behaviors and preferences of users engaged in exploratory searching, the tasks supported by ESSs, and on measures of exploration success. While search systems are expanding beyond the support of simple lookup into the support of complex information-seeking behaviors, evaluation of search systems has remained limited to those that encourage minimal human-machine interaction. Information Retrieval (IR) is by nature an experimental discipline; the evaluation of retrieval algorithms and other aspects of system design such as document indexing and the user interface are central to the progress of the field. The Cranfield methodology (Cleverdon et al., 1966), that later became utilized by the NIST-sponsored Text Retrieval Conference (TREC) (Harman, 1993), has been a useful paradigm for the objective comparison of IR systems, where only one aspect of a system was varied at any point in time. TREC has provided a medium for the evaluation of algorithms underlying the analytic aspects of IR systems, yet struggled because the experimental methods of batch retrieval are not suited to studies of how search systems are used by human searchers. Search systems are not used in isolation from their surrounding context, i.e., they are used by real people who are influenced by environmental and situational constraints such as their current task. To be used effectively search systems must have provision to adapt to these contextual constraints (Ingwersen & Järvelin, 2005), and therefore evaluation methodologies must be able to evaluate systems on this basis. Since TREC-3, the conference has extended its mandate to recognize the importance of the user in information-seeking. The Interactive Track (Dumais & Belkin, 2005), and later the High Accuracy Retrieval of Documents (HARD) track (Allan, 2003) 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 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 information-seeking (Lagergren & Over, 1996). Whether or not the evaluation of exploratory search will blossom within the TREC paradigm remains to be seen; however, it is clear that researchers are increasingly turning their attention toward new ways to systematically investigate the effectiveness of ESSs and the more general information seeking process.

Exploratory search systems capitalize on new technological capabilities and interface paradigms that facilitate an increased level of interaction with information. High levels of interaction, which are an integral part of exploratory search, pose a real evaluation challenge: there is potential for confounding effects from different exploration tools, the desired learning effect is difficult to measure, and the potential effect of fatigue limits the evaluation to a low number of topics, which makes it rather difficult to get the statistical significance required by a meaningful quantitative analysis. A key component of exploration is human learning, a topic studied extensively by cognitive psychologists (Landauer, 2002). Indeed, subject-matter learning has been proposed as a way to evaluate exploratory search systems as a function of exploration time and effort expended. Support for more-rapid learning across a number of users and a range of tasks is indicative of a system that is more effective at supporting exploratory search activities. For example, in an evaluation of Scatter/Gather, an interface designed to support search result exploration through text clustering, Pirolli and colleagues (1996) attempted to measure learning and understanding in terms of

topic structure and query formulation capabilities at various points during subject interaction with the system. In comparison to a control group that did the same tasks using a standard search engine, users of the Scatter/Gather system showed larger gains in understanding the underlying topic structure and in formulating effective queries. The similarities between exploratory search and sensemaking / information foraging mean that an analysis of the costs involved in the process in terms of gain for time spent representing / understanding the task and finding / selecting information may also be useful for comparing exploratory search systems (Russell et al., 1992). Ultimately of course, we must measure the depth and effectiveness of learning rather than focusing on efficiency and thus time may be less appropriate as an outcome measure. Nonetheless, time to learn may be an excellent metric of choice for exploratory search at this stage in the development of ESSs.

Evaluating ESS is not substantially different from evaluating any other highly interactive system. Whilst of course we should be concerned with subjective measures such as user satisfaction, engagement, information novelty, and task outcomes, it is through the measurement of interaction behaviors, cognitive load, and learning that we can get a clear picture of how effective are ESSs. There are research opportunities to develop frameworks for the evaluation of ESS that incorporate such measures. The approach adopted at TREC has led to the rapid development of effective ranking algorithms for document retrieval. As a result of such research, search systems such as Google, Yahoo!, and Live Search cope well with navigational requests (e.g., find a specific person's homepage), and closed informational requests (e.g., answer to a question which has a single answer). However, none of these systems provides the explicit functionality to support exploration. It has been suggested that repositories of data and tasks (similar to TREC) could be used to evaluate ESS based on information visualization (Plaisant, 2004).

Our long-term vision is for a framework for ESS evaluation that could validate the support these systems offer, and chart new courses toward improved search experiences for users. The articles in this special topic issue take an initial step in this direction. They discuss issues related to the formative and summative evaluation of systems that support exploratory search activities, as well as evaluation methodologies and paradigms, and describe the practical evaluation of existing ESSs.

Ruthven and colleagues investigate how document surrogates might be useful in exploratory search and what information it is useful for a surrogate to contain. Through comparing assessments based on artificially created information surrogates, they investigate the effect of the source of information, the quality of an information source and the date of information upon the assessment process. In addition, they also investigate how varying levels of topical knowledge, assessor confidence and prior expectation affect the assessment of information surrogates.

Kules and Shneiderman advocate categorized overviews of Web search results – combining a metadata-based overview with search results – to support user exploration, understanding, and discovery. They describe and present results from a large mixed-methods study of sophisticated users carrying out complex tasks using categorized overviews of Web search results organized into thematic, geographic, and government categories. The article

includes a qualitative analysis of searcher comments that identified tactics that participants reported adopting when using categorized overviews and a set of guidelines for the design of exploratory search interfaces.

Hoeber and Yang describe an evaluation of the "WordBars" system that assists users in their Web search and exploration tasks, in particular with query refinement activities. The system provides a visual depiction of the frequencies of the terms found in the top-ranked document surrogates returned from an initial query, in the form of a histogram. Exploration of the search results is supported through term selection in the histogram, resulting in a re-sorting of the search results based on the use of the selected terms. User evaluations with both expert and intermediate Web searchers illustrate the benefits of the interactive exploration features of WordBars in terms of effectiveness as well as subjective measures.

He and colleagues focus on task-based information exploration by intelligence analysts in performing tasks that are complex, dynamic, multi-faceted, and multi-stage. They present an evaluation framework designed specifically for assessing and comparing performance of information access tools at analysts' major information access stages, such as information foraging and sensemaking. The framework is accompanied with a reference test collection that has a number of tasks scenarios and corresponding passage-level ground truth annotations. To demonstrate the usage of the framework and the test collection, they present a specific evaluation study on CAFÉ, an adaptive filtering engine designed for supporting task-based information exploration.

Qu and Furnas present a model-driven formative evaluation of basic exploratory search systems in the context of a sensemaking task. Study participants were asked to make sense of an unfamiliar topic using an augmented query-based search system. The processes of representation construction and information seeking were captured and analyzed using data from experiment notes, interviews, and a system log. The data analysis revealed sources of ideas for structuring representations and a tightly-coupled relationship between search and representation construction in exploratory searches. The authors claim that subjects searched to find useful structure ideas instead of just accumulating information facts. In addition they present the design implications of their study.

An earlier special section of Communications of the ACM entitled "Supporting Exploratory Search" (White et al., 2006) focused on the design of interfaces to help users conduct exploratory searches. That collection of papers has been successful at highlighting the exploratory search challenge and in driving the development of interfaces designed to address it. In this special topic issue we continue to encourage work on this important problem, and highlight research being conducted in the area of *evaluating* exploratory search systems. As can be seen from the brief descriptions in this introduction we have selected articles that cover a range of aspects of evaluation. The objective is to showcase some research in ESS evaluation, and bring the challenges and opportunities of evaluating systems that support users engaged in exploratory search scenarios and adopting exploratory search strategies to the attention of the research community. It is our hope that the articles presented herein inspire others to focus attention on these and similar areas.

References

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

Cleverdon, C.W., Mills, J., and Keen, M. (1966). *Factors determining the performance of indexing systems*. ASLIB Cranfield project, Cranfield.

Dervin, B. (1998). Sense-Making theory and practice: An overview of user interests in knowledge seeking and use. *Journal of Knowledge Management*, 2(2): 36-46.

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.

Harman, D. (1993). Overview of the First Text Retrieval Conference. In *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 36-47.

Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer.

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.

Landauer, T.K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *The Psychology of Learning and Motivation*, 41: 43-84.

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4): 41-46.

Marchionini, G. and Shneiderman, B. (1988). Finding facts vs. browsing knowledge in hypertext systems. *IEEE Computer*, 21(1): 70-79.

Pirolli, P.L. and Card, S.K. (1995). Information foraging in information access environments. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 51-58.

Pirolli, P., Schank, P., Hearst, M., and Diehl, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 213-220.

Plaisant, C. (2004). The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pp. 109-116.

Russell, D. M., Stefik, M. J., Pirolli, P. L., and Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 269-276.

Shneiderman, B. and Plaisant, C. (2004). *Designing the User Interface 4th Ed.*, Pearson/Addison-Wesley.

White, R.W., Kules, B., Drucker, S., and schraefel, m.c. (2006). Supporting exploratory search: Introduction to special section, *Communications of the ACM*, 49(4): 36-39.