

Evaluating Exploratory Search Systems

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

Gheorghe Muresan
School of Communication,
Information and Library Studies
Rutgers University
New Brunswick, NJ 08901 USA
muresan@scils.rutgers.edu

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

Online search has become an increasingly important part of the everyday lives of most computer users. Generally, popular search tools support users well, however, in situations where the search problem is poorly defined, or the information seeker is unfamiliar with the problem domain, or the search task requires some exploration or the consideration of multiple perspectives, such tools may not operate as effectively. To address situations where technology may not meet their needs, users have developed coping strategies involving the submission of multiple queries and the interactive exploration of the retrieved document space, selectively following links and passively obtaining cues about where their next steps lie. This is an example of *exploratory search* behavior, and comprises a mixture of serendipity, learning, and investigation [7].

Exploratory search can be seen as a specialization of information exploration – a broader class of activities where new information is sought in a defined conceptual area. It represents a shift from the analytic approach of query-document matching toward direct guidance at all stages of the information-seeking process. Through functionalities such as tabbed browsing and dynamic queries [10], Exploratory Search Systems (ESS) are helping users run multiple threads in parallel, and see the immediate impact of their decisions. By following hyperlinks, people can better define and refine their information problem, and bring it closer to resolution. Browsing is a serendipitous activity that can be attractive to users, who may benefit from the extraneous information if they have long-term interest in a particular topic, but is inefficient for fact-finding or known-item retrieval, so is therefore not appropriate for all circumstances [8].

Browsing a new document collection, beginning to gather information on a new topic, or trying to resolve an ill-defined problem, can be likened to the exploration of a maze in the physical world; the process is fraught with uncertainty, one is never able to see more than one step ahead at any given time, and the navigation of the maze comprises a series of on-the-fly selections that can impact the success of the journey. Analytical strategies that provide us with a ranked list of documents can be seen as providing a point or entry to the maze, or even dropping us in the middle. However, to find the prize at the center of the maze (or escape from it!) there is a need to provide tools to support navigation and decision-making. For example, finding one's way through the maze becomes much easier if a visual representation of the space being explored is provided (e.g., map

with current location indicated). Now, imagine that the maze is multi-dimensional, and that choices at each intersection are not limited to one out of two, three, or four possibilities, but rather tens and hundreds of possibilities, as is the case with exploring search results. The design of interfaces to help users navigate these complex environments is a crucial part of supporting exploratory search, and outweighs the analytic strategies prevalent in current search systems, which serve to parachute us into a starting point. As the articles in a recent issue of *Communications of the ACM* entitled "Supporting Exploratory Search" [11] demonstrate, research into the development of interfaces to support the understanding of information, rather than simply finding it, is gathering pace in communities such as human-computer interaction, information retrieval, library and information science, psychology, and beyond.

Exploratory search systems are capitalizing on new technological capabilities and interface paradigms that facilitate an increased level of interaction with information. However, evaluation of search systems has remained limited to those that support minimal human-machine interaction. Since the days of the Cranfield experiments some 40 years ago, the issue of evaluating retrieval systems has been considered highly important by the Information Retrieval (IR) community [2]. The annual NIST-sponsored Text Retrieval Conference (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 interactive IR. Since TREC-3, the conference has extended its mandate to recognize the importance of the user in information-seeking. The Interactive Track [3], 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 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 [5].

The more interactive options an application has, the greater the number of variables, and therefore the larger the likelihood for experimental confounds if compared against other systems. For example, a system with features A, B, C, D, and E should theoretically be compared against 119 other systems that vary the presence and absence of these five features. Even if the experimenters make pragmatic decisions about the number of experimental variations, it is still challenging to limit the number of comparator systems whilst maintaining control of the number of possible experimental confounds. This does not include the time required to complete the experiments, build the systems, and train the subjects using the systems.

Additionally, it may often be the case that the sum of the features in an ESS may lead to a different experience than the individual features in isolation (i.e., the interactions between features may be just as important as the features themselves). High levels of interaction, which are an integral part of exploratory search, pose a real evaluation challenge: there is potential for confounding effect of 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.

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 is 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 [6]) yet this issue has not been explored in relation to ESS. Any evaluation of ESS should consider at least two factors: *metrics* (i.e., what is going to be measured?), and *methodologies* (i.e., how we are going to measure it?).

Metrics: The outcomes of the search and the search process itself can be used to evaluate the effectiveness of ESS. For example, assessments of relevance or utility by subjects during or after the search, structured or informal subjective evaluations, and examination of the resultant products or artifacts, all give insight into the effectiveness of the ESS. However, they give limited insight into how well systems support cognitive processes such as learning. One way to get access to such information is to look at users' interactions during their search. Behaviors can be seen as manifestations of internal information-seeking strategies. An examination of paths taken and decisions made during a search can allow us to make inferences about cognitive activity [8].

Methodologies: The approach taken to evaluate ESS is crucial. If possible, experiments should be longitudinal, and take place in a naturalistic setting. The task domain should contain a mixture of task types: some that relate closely to subjects' regular activities, and some that are completely new. A challenge of ESS evaluation is to elicit exploration, and this can be more problematic if subjects are only engaged in tasks they are familiar with. Subjects should be classified based on familiarity with the topic or problem domain, expertise or frequency of using the retrieval system, and general range of computer experience. The setting and task domain should be controlled by the experimenter, to allow focus on the user and the system components of information-seeking. To counteract learning or order effects that may compromise the reliability of the experimental findings, there should be systematic variation of the independent variables in the experiment. Exploratory search is a cognitively intensive activity, and subjects should be allowed to conduct their searches with minimal interruptions. Techniques such as questionnaires and interview techniques can be valuable tools, but one must be careful to include them in the experiment in such a way as not to interfere with their exploration. If multiple sites are going to be involved in the experiment then care should be taken to coordinate planning and execution carefully.

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 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 such systems can be. 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 MSN Search and Yahoo! cope well with navigational requests (e.g., find a given 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 [9]. 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.

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] Cleverdon, C.W., Mills, J., and Keen, M. (1966). *Factors determining the performance of indexing systems*. ASLIB Cranfield project, Cranfield.
- [3] 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.
- [4] Kelly, D. (2004). *Understanding implicit feedback and document preference: A naturalistic user study*. Unpublished doctoral dissertation, Rutgers University.
- [5] 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.
- [6] Landauer, T.K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *The psychology of learning and motivation*, 41: 43-84.
- [7] Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4): 41-46.
- [8] Marchionini, G. and Shneiderman, B. (1988). Finding facts vs. browsing knowledge in hypertext systems. *IEEE Computer*, 21(1): 70-79.
- [9] Plaisant, C. (2004). The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced Visual Interfaces*, pp. 109-116.
- [10] Shneiderman, B. and Plaisant, C. (2005). *Designing the User Interface 4th Ed.*, Person/Addison-Wesley.
- [11] White, R.W., Kules, B., Drucker, S., and Schraefel, M.C. (2006). Supporting exploratory search: Introduction. *Communications of the ACM*, 49(4): 36-39.