Evaluating IR in Situ

Susan Dumais
Microsoft Research, Redmond

Information retrieval has a long and successful tradition of careful evaluation using shared testbeds of documents, queries, relevance assessments, and outcome measures. This paradigm has served us well for improving representations, matching and ranking algorithms, but it has limitations. Evaluations methodologies need to be extended to handle the scale, diversity, and user interaction that characterize information systems today. Previous research on interactive IR has focused on small-scale laboratory experiments. In contrast, Web search engines, e-commerce sites, and digital libraries all benefit tremendously from being able to study large numbers of searchers in situ as they interact with information resources using log data and/or more controlled experiments. Such data provide valuable insights about what users are doing, and how well current search systems are meeting those needs. There are important challenges in collecting and using interaction data (e.g., privacy of individual data, replicability of experiments in the face of changing content and queries, extracting signal from noisy behavioral data), but I believe that we must begin to address these issues and extend our evaluation methods to make continued progress in IR.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.