Supporting interaction and familiarity

James Allan, University of Massachusetts Amherst
Information Seeking Support Systems Workshop, June 26-27, 2008
Position Paper – Draft for workshop

I am interested in building "information seeking" systems to support expert searchers (not the general populace) in the tasks of collecting, organizing, and monitoring information related to their issues of interest. Examples of expert searchers include political analysts tracking an important issue, researchers checking on the latest developments in some experiments, or information analysts watching for terrorist activities in Europe. Rather than document retrieval, I am focusing on "fact finding", more along the line of several Question Answer tasks, including "complex interactive QA" that was recently run at TREC.

The particular inspiration for this position paper comes from a series of studies carried out by Patterson, Woods, and Roth at the Ohio State University's Institute for Ergonomics, in the Cognitive Systems Laboratory. (I make no claim that this is the only study of this sort; it is merely the immediate inspiration of this paper.) They invested considerable effort in understanding how people, particularly information analysts, cope with problems of data overload. In exploring how humans extract useful information from data, their work points toward some important criteria that a support system needs to satisfy.[1] Based on their observations, I propose two ideas as important for information seeking support systems: (1) assisted information finding and (2) moving from familiar to unknown sources.

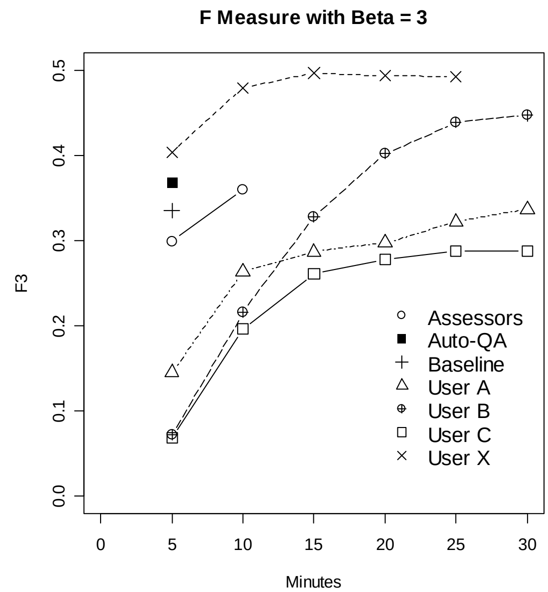
Assisted information finding

The expert searcher I aim to support wants to find a collection of facts and relationships that answers a question. The question is complex, perhaps even difficult to articulate succinctly—certainly not in a 2½-inch field on a Web page. The searcher may know in advance whether there are likely to be a few or hundreds of useful facts, whether some or all are needed, and whether the facts are static or constantly being expanded upon. In such an environment, the classic model of information finding no longer makes sense: an effective system cannot accept a small query, list documents in response, and then wait to see if any are selected. A system designed for this situation must engage with the searcher, working with him or her to develop an expressive query, to extract useful facts, to identify their relationships, and then build incrementally on the results by repeating the process if necessary. (This is probably *Motherhood and Apple Pie* to the attendees of the workshop.)

Existing search and organization systems tend toward one of two extremes. The most common approach is the simple search engine, exemplified by all major Web search engines. They accept a query and present a list of documents ranked by the likelihood of relevance. They might remember and be able to leverage past queries, and might be able to recommend documents that were interesting to

[1] D.D. Woods, E.S. Patterson, and E.M. Roth (2002), "Can we ever escape from Data Overload? A cognitive systems diagnosis." *Cognition, Technology, and Work*, Springer London, 4(1):22-36.

other people. They do not work in conjunction with the searcher. At the other extreme are powerful visions of systems that do everything for the expert: given a query, they are intended to return precisely the desired list of facts. These systems are the goals of government funded initiatives such as DARPA TIDES, DARPA GALE, and the ARDA/DTO AQUAINT programs. The most successful work in those areas can answer simplistic questions well and complex questions with widely varying but limited success.



We have recently developed and evaluated (as part of TREC 2007's ciQA track) an interactive search system designed for finding answers. The system allows the searcher to describe a complex information need (though participants only used free text descriptions), to see documents that were on the right topic, and then to identify answers to the need. The graph on the right shows search effectiveness on the vertical axis and time on the horizontal. The best performance (across the entire evaluation), regardless of time on task, was achieved by an expert searcher ("User X") familiar with the system. Using a system designed to help a trained searcher find answers to questions can result in substantial gains over purely automatic techniques. However, major challenges remain.

Two key challenges are helping the searcher develop a framework for organizing his or her thoughts and using that framework both to help organize results and to help formulate queries. When the expert searcher knows what sort of result is likely to be correct in advance, the search approach can be tuned to expand or restrict the scope of the search, to look for supporting evidence or counter-examples, and to develop queries based on positive or negative examples. Our prototype system provides a rudimentary version of the needed framework, but does not use it to assist with the actual search process. I suspect that relevance feedback may be able to leverage negative information more usefully in such an environment.

From the familiar to the unknown

The 1999 study by Patterson et al. shows a common practice of analysts faced with daunting amounts of raw data. When trying to find their answers, they tend to "narrow returned sets based on the number of hits almost indiscriminately…." [2] We and others in the research community have similarly found that searchers prefer a sense of control, whether it is of the data being explored or of the language used to describe requests. Most implemented search systems address this issue in limited ways, if they do at all. They might provide very simple search capabilities to avoid overwhelming the searcher—e.g., Web search engines allow keywords with a handful of rarely used operators—or they might construct

[2] E.S. Patterson, E.M. Roth, and D.D. Woods (1999), "Aiding the Intelligence Analyst in situations of data overload: a simulation study of computer-supported inferential analysis under data overload." Interim report to the Air Force Research Laboratory, AFRL-HE-WP-TR-1999-0241, p. 77.

elaborate visualizations intended to provide better understanding, hence control—e.g., color bars used to identify which query terms appear in a document and where. Neither approach actually recognizes that searchers are more comfortable with smaller amounts of data.

I believe that a key challenge for information seeking support systems is developing techniques and interfaces that allow a searcher to work in three stages: (1) facilitate rapid and appropriate winnowing of material to a small and comfortable set of items, (2) support exploration and organization on the small set, and (3) enable expansion back to the original large data set to find information that supports, refutes, or builds on what was found in the small set. Each of those stages requires investigation of techniques that work for searchers and algorithms that support the data processing efficiently and effectively.