

Measuring the Tools and Behaviors of Sensemaking

Daniel M. Russell

IBM Almaden USER Lab
650 Harry Rd., San Jose, CA
daniel2@us.ibm.com

Malcolm Slaney

IBM Almaden USER Lab
650 Harry Rd., San Jose, CA
malcolm@ieee.org

ABSTRACT

We present a method for assessing one kind of sensemaking behavior — rapid understanding of large document collections — and discuss lessons learned from our attempts to measure representative sensemaking tools and sensemaking behavior. To further our understanding of sensemaking behaviors, we need to move beyond overly constrained and artificial measurements of easily instrumented behavior. From observations, we know sensemaking is often performed under time pressure and requiring use of large document collections. Instrumenting people in their workplace is often untenable, yet oversimplified laboratory studies often miss explanatory richness. We argue that studies of sensemakers need to be done on tests that are closely aligned with the natural tasks of sensemakers. Understanding human performance in such tasks requires analysis that accounts for many of the subtle factors that influence final performance, including the role of background knowledge, variations in reading speed, and tool use costs.

Author Keywords

Sensemaking; document collections; visualization; measuring analysis work; information retrieval.

ACM Classification Keywords

H.5.2. user interfaces and presentation; H.3.3: Information search and retrieval; H.1.2 User/Machine Systems.

SENSEMAKING IN LARGE COLLECTIONS

How do people understand their complex, information-rich world? *Sensemaking* is the process of a person coming to understand a large body of information. Here, we look at an aspect of the sensemaking problem, specifically, how well different tools can help someone understand and build a good mental model of the information in a large document collection. Our goal is to study how people perform on the entire task.

KEEP THIS SECTION CLEAR.....

Information retrieval (IR) is a well-studied area, with well-understood means of performance assessment. The most common performance-measuring approach is to evaluate the precision and recall of the results from a single search. Clustering and relevance sorting are often used to further organize the results [6]. By contrast, sensemaking is the entire set of behaviors involved in coming to a deep understanding of a body of content, rather than just measuring the effectiveness of a search in isolation.

In the information foraging analytical framework, information analysts are characterized as sometimes searching for information, deciding how much time to put into one document, before deciding to move on, and selecting a new source of information [4]. The infoscent model does a good job of predicting which web link a subject will click on next. But again, this work assumes a single reasonably-formed goal (e.g., find a picture of a performer for an upcoming concert) that can be evaluated at multiple points during the search, as well as a working style that is primarily link-following, rather than collection understanding.

By contrast with these common characterizations, sensemaking users are frequently presented with problems that do not fit into the narrow IR or information-foraging categories [1]. Quite often, sensemakers are handed a large collection of documents and need to understand it in a general way, often to write a critical analysis or report. “And, by the way, can you please have an answer in 1 hour?” Real sensemaking is often done under extreme time and performance pressure [11].

Part of the sensemaking model is information modeling and synthesis. In fact, our tasks are defined in terms of how well a human subject understands the material. As described in earlier work [5], each portion of the sensemaking process comes with a real cost. In Figure 1, this depiction of sensemaking shows that subjects must form a model, organize the information they have into facets of this model, and most importantly, update the model, perhaps even throwing it out, when the facts processed by the user no longer fit the model.

Our work here steps back and looks at the bigger picture. How well can human subjects search, study, iterate and integrate the information from a large collection of documents? We want to measure user performance on the

entire problem, including aspects of searching for a representation, finding instances of it, and then using the results of sensemaking in a complex, realistic task.

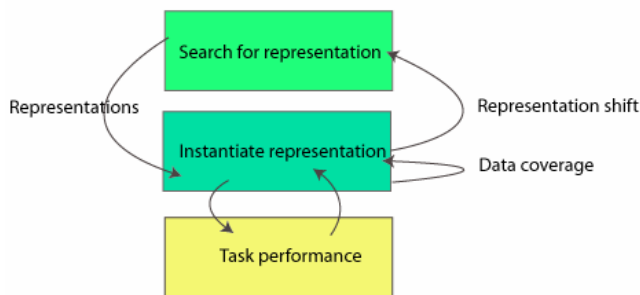


Figure 1: Sensemaking is an iterative process of reading information, creating representations of it, assessing the goodness of the model and content created, then using the content in a task.

This approach is distinct from question answering (e.g., [10]) since sensemakers do not know the specific questions they will have to answer ahead of time. They *are* on an information-foraging expedition, one that requires learning as much as they can so that they come to understand the issues in the document collection. In many ways, our task is similar to a document-summarization problem. Our subjects must understand the gist of a collection, so they can answer questions about it.

MEASURING SENSEMAKING BEHAVIORS

The aspect of sensemaking tested here is not simple. When users come to a broad understanding of a document collection, there is no immediately obvious goal when they begin to work. Instead, the more the subject reads through a collection, the more organized their knowledge about the domain becomes. It is unclear exactly what the subject's goal should be, other than to generally understand in anticipation of future use.

Despite a wide interest in questions of how people actually use information retrieval tools in sensemaking tasks, there is a surprising lack of attention paid to details of realistic uses.

For example, only rarely does search, per se, enter into the problem. In some studies of sensemaking behaviors, actual *search time* is not the majority of the total time on real tasks [3]. Instead, users have many goals, only some of which are typical search goals. And for many tasks, goals evolve as more is learned about the subject area and the actual output of the work process [7]. Thus, seldom can sensemaking behavior be characterized as an information retrieval problem, addressed by one query and ideally served by a single retrieved document. Although perfect precision is a laudable goal, given the shifting nature of sensemaking goals, it is not clear that perfect precision would improve the user's overall performance.

Real sensemaking behavior has features not taken into account with many information retrieval assessment

models. Sensemakers (1) often begin a task with a goal that's well-defined, but untranslatable into IR actions at the outset, (2) when working with a search tool, they often reformulate queries as they work towards creating a finished product, and (3) bring a great deal of background knowledge to the task (including things they know/recognize in context, but don't know they know).

Thus, we need to focus on how analysts perform in these tasks. It is often the case that even low-precision tools are amazingly useful in the hands of an experienced analyst. The very notion of precision and recall measures are open to question with a source collection (e.g., the web or large proprietary repositories) that change enormously moment-to-moment and in the face of queries that evolve in midstream.

Our initial goal was simply to measure the effects of different common information visualizations. But when we began doing pilot tests in this area, it became clear that devising an adequate test that takes into account realistic sensemaking behaviors required a more complete test framework.

CREATING A TEST TO MEASURE SENSEMAKING

We want to understand how people perform in realistic sensemaking tasks, using real data and tools. In our experimental paradigm we measure the end-to-end performance in a typical intelligence analytic task.

Our test measures how well subjects understand a large corpus of documents when given different amounts of time. This task is commonplace in many intelligence and analysis work settings, and, as such, offers a realistic (if complex) way to measure the effects of search and visualization tools. The test is sensitive to differences in the tools, and as we discovered, such tests also reveal substantial individual differences in performance. In this paper we explore the issues involved in such a test. In a companion paper [10] we show detailed results for a set of experiments involving 24 subjects and three different tools.

Understanding large document collections (ULDC): The ULDC test measures the ability of subjects to understand a large document collection after spending time examining it. Subjects spend varying amounts of time in examining the collection (where they can skim, scan, read, take notes, etc.), and then are given a multiple choice test to measure their level of integrated understanding of the collection.

In designing the ULDC test, we chose a collection of newspaper articles from cities that we believed were relatively obscure in order to put all subjects onto a equal footing in their lack of background knowledge. We briefly considered using articles from a technical field, such as chemistry, but knew we would have a hard time finding subjects with sufficient background knowledge to understand the corpus. By contrast, all our subjects have a good ability to understand the documents in a news collection.

For this test we set up six document collections, each containing 300 articles about a city drawn from the LDC GigaNews corpus. [2] The cities (Vladivostok, Paramariba, Katmandu, Baku, Bilbao, Riga) were chosen because they had a good number of articles about a handful of issues and relatively little sports news (which were mostly rather brief news articles with scores, not amenable to sensemaking efforts.) The 300 longest articles from each were chosen for the test collections.

Subjects create models of each collection in their heads, sometimes taking notes of their choosing, writing down important information on a pad of paper. We measure the subject’s success at the sensemaking problem by testing the quality of their mental model through a questionnaire given them after a sensemaking session of either 0, 5 or 15 minutes duration.

Test creation: A professional news editor read the paper version of each document collection in detail (taking as much time as necessary, typically several hours) creating 30–40 multiple choice questions to test a subject’s *integrated understanding* of the important news in each city. That is, the questions were intentionally designed to require information that needed some integrating thought, rather than simple fact retrieval. A sample question: “What were the main three issues in the Vladivostok mayoral election of 1998?” (There is no single article in the collection that explicitly lists the answer to this question.)

Measuring tool variation: We wanted to understand how different tools affected subject’s abilities to understand the collection of documents. Subjects were tested on three different access methods for the documents: a bound volume with paper printouts of the articles from each city and two different electronic visualizations. The “semantic tool” presented users with a fairly standard scatter plot of documents, organized by LSI dimension reduction, which gives a 2D distribution of points, roughly organized into semantically meaningful regions of the display, such as those shown in [8]. The “temporal tool” showed a time-line distribution of the documents. In both tools, rolling the mouse over the document icon gave a tooltip-style summary of the document, while clicking on the icon brought up a document window with the full document text for close reading.

RESULTS OF ULDC TEST ON 6 CITY COLLECTIONS

Figure 2 shows the performance for 12 subjects who took each test of city knowledge under each of our 3 conditions (0 minutes of study, aka “baseline”; 5 minutes of study; and 15 minutes of study). The 3 curves (each a Gaussian approximation to the raw data) represent each of the three different visual presentations (paper document, temporal, semantic). Each subject took all six tests (one / city) and the results show the average number of questions answered correctly. As expected, with more time for study of a collection, subjects could perform better on the tests.

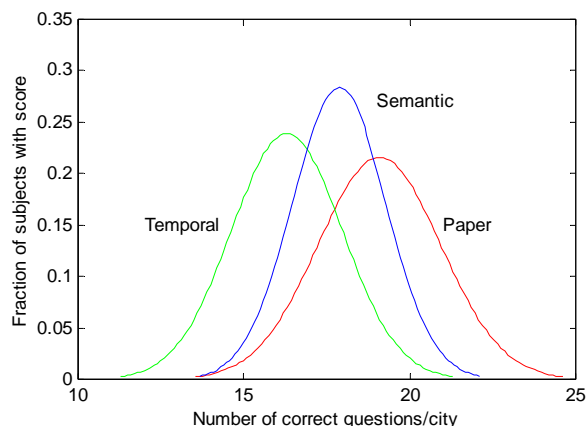


Figure 2: Subject performance improves with total time on task, regardless of tool used to inspect the document collection. A curve farther to the right indicates better test performance. The left-most curve is for 0 minutes (i.e., baseline), the middle curve is for 5 minutes of study, and the rightmost is after 15 minutes.

But as Figure 3 shows, somewhat surprisingly, subjects who used one of the two information visualizations to study a city’s documents performed poorly, compared to those that used the paper collection.

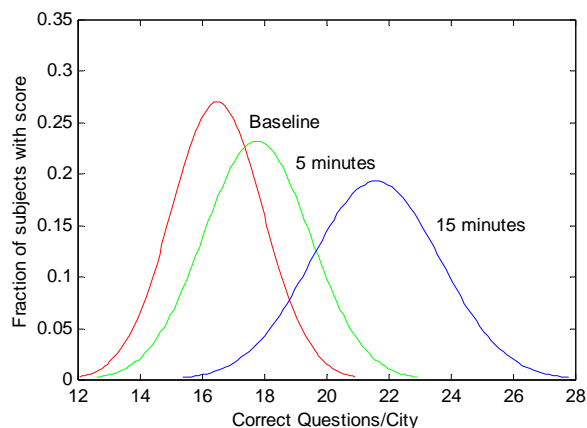


Figure 3: Each visual document collection examination tool gives the user a different level of understanding. Here, a bound paper collection outperforms both a temporally-organized and semantically-organized visualizations. Curves to the right indicate more correct answers. The curves are significantly different.

LESSONS LEARNED

In the process of creating the ULDC test, we found that measuring sensemaking behaviors for a time-paced, integrative, collection-understanding task is full of subtleties. There are many inherent sources of variance or noise in studies of this kind, factors that need to be taken into account whenever a user study of sensemaking is undertaken.

Tools matter: Our biggest surprise was that our attempt to give subjects fairly standard visualization tools to help their

performance actually hindered them. Tools can help, or be problematic: measurement counts, intuitions do not.

1. Background knowledge: There is apparently no realistically-testable topic on which we can measure subjects without some useful background knowledge. Some people might be mountain climbers and be especially interested in news from Katmandu. In addition, all people will have different specific knowledge that will bias their internal models. This prior knowledge includes subtle knowledge the subject knows and recognizes in context, but don't know they know—all of which serves to help the subject organize the material they are studying. But different people will have different knowledge, and almost by definition, our subject pool of researchers is broad and deep—this leads to a variation in scores. Yet, this diversity seems common among the knowledge worker population of sensemakers. We need to accept the inherent variance in human backgrounds and abilities, and work to characterize it as part of the inherent measurements.

Baseline studies: The baseline cases (n=12) were of subjects who took the test about each city without any exposure to the document collections. In essence, this pool of subjects acts as our control group for comparison purposes. Although we purposefully chose domains that were obscure, there was still a wide variation in the outcomes. But doing this baseline data collection was important for our comparisons with experimentally conditioned data. We could not have understood the data without this baseline data as a reference point.

Reading speed dramatically influences outcomes: An earlier pilot experiment, with just six subjects, demonstrated the subtleties involved in the choice of test subjects. Subjects for both the baseline experiment and for the visualization tests were chosen from a population of full-time professional researchers at our laboratory site (only 4 were from our lab). We were surprised that some subjects took significantly longer to complete the test than others. Each subject had three 15-minute tests, and three 5-minute tests, so total time for reading documents was held to 60 minutes. Yet we observed some subjects taking much more than 2 hours to read the documents *and* complete the test questionnaire.

Later, when studying the composite scores, we realized that some subjects, while excellent researchers and speakers of English, did not learn English as a first language. On further reflection and discussion with our non-native English colleagues, we found that reading speed is often much slower for non-native speakers. Skimming speeds, a skill necessary for our time-paced, rapid sensemaking task, can be as much as an order-of-magnitude slower.

SUMMARY

We have presented here our sensemaking test paradigm for large document collections, highlighting the issues that arise during testing and analysis.

Implications of the study: (1) It is important to measure consistent subject populations, rather than trying to post-hoc analyze the data to reduce possible sources of variation (that is, operatively, we attempt to reduce variation by careful selection of subjects to study in depth, rather than having a broad population.) (2) We found it important to track native vs. non-native speakers (because of different reading styles). Non-native readers perform significantly differently in skimming and scanning skills, which are central skills in this paradigm. (3) Control baseline study carefully by testing with subjects as closely matched to experimental subjects as possible. (4) Finding obscure content material to limit the effects of background knowledge is difficult, if not impossible. Sensemaking studies must account for these effects in their analysis. (5) Test questions need to be written to test for integrative understanding, not simple fact retrieval. (6) Sensemaking testing must be conducted with a significant number of documents (to stress human limits, not just play to strengths of tool under study).

REFERENCES

1. Heuer, R. J. Jr. *Psychology of Intelligence Analysis*. United States Government Printing Office, November 1999.
2. Linguistic Data Consortium. The GigaWord Corpus. <http://www ldc.upenn.edu/Catalog>
3. O'Day, V., Jeffries, R. Orienteering in an information landscape: How information seekers get from here to there. Proceedings of InterCHI '93, Amsterdam, Netherlands, 1993.
4. Pirolli, P., Card, S., Information foraging. *Psychological Review*, 106: 643–675, 1999
5. Russell, D. M., Stefik, M. J., Pirolli, P. and Card, S. K. The cost structure of sensemaking. Proceedings of InterCHI '93, Amsterdam, Netherlands, 1993.
6. Salton, G., & McGill, M. *Introduction to Modern Information Retrieval*, New York, McGraw Hill, 1983
7. Spink, A., Jansen, J., Ozmultu H. Information seeking and mediated searching study. Part 3: successive searching, *Am. Soc. Inf. Sci. Technology*, 53 (9) pp. 716-727, 2002.
8. Wise, J. A. The ecological approach to text visualization. *Journal of the American Society for Information Science* 50 (13): 1224-1233, 1999.
9. Voorhees, E. M. Overview of the TREC 2002 question answering track. *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*, 2003.
10. Slaney, M., Russell, D. M. How to grok a collection: the effectiveness of information displays in a multi-document collection, Proc. CHI 2004 (in progress).
11. Patterson, E. Computer-supported inferential analysis under data overload. *Proc. CHI 1999*, Pittsburgh, PA, 2002.

