Being literate with large document collections: Observational studies and cost structure tradeoffs

Daniel M. Russell[1], *Malcolm Slaney*[2], *Yan Qu*[3], *Mave Houston*
IBM Almaden Research Center
650 Harry Rd.
San Jose, CA 95120, USA
daniel2@us.ibm.com, malcolm@ieee.org, yqu@umich.edu, maveh@us.ibm.com

Abstract

How do people work with large document collections? We studied the effects of different kinds of analysis tools on the behavior of people doing rapid large-volume data assessment, analysis and organization. We analyzed the micro-structure details of using automated clustering techniques, the use of standard timeline and cluster visualization methods, alongside desktop paper sorting and piling. We find that the "natural" methods people use (with piles of paper documents) are in fact very sophisticated and have a subtlety that is lacking in current computer interfaces. This analysis shows that the lack of expressiveness and responsiveness in current interface designs dramatically limits human performance, suggesting ways in which the next generation of analytic tools must evolve in order to support literate use of large volume / complex document collections.

1. Introduction

"Making sense" of a large document collection—what we call *sensemaking*—can be seen as the process of creating a representation of a large volume of information that allows the analyst to perceive structure, form and content within a given corpus. It's what people naturally do when faced with too much information to handle in a short amount of time. This kind of large corpus understanding is a fairly typical analysis task in a number of intelligence and business settings. We are especially interested in what people do when faced with sensemaking tasks that use large document collections: such tasks seem to be central to many kinds of intelligence analysis tasks in governmental, business and personal domains. [16]

When people need to rapidly make sense of a large document collection they usually begin by skimming the documents and organizing the collection into temporary groups (clusters). This sensemaking behavior gives a quick overview of the contents, while creating a fast, easy to use representation for organizing and accessing the accumulated contents. In this study we contrast the time and effort subjects put into sensemaking of document collections for both manual manipulations of physical documents and when using online clustering tools. Our close analysis of how people spend their time when working with the documents, the representations they make, and the way they choose to spend their time defines our cost structure model of analyst behavior. The characteristics we find in the underlying model suggest that basic task structure of human sensemaking is very sensitive to the design and costs of using online tools.

A common assumption is that almost any kind of automated assistance will improve human performance. We began this series of studies with this naïve assumption as well. As we found out, in many cases, tools that are not well-designed to match human capabilities can actually slow down performance, particularly when used in stressful conditions. We expected visualization tools and automatic clustering to help, but they didn't in every case. Why?

[1] Now at Google
[2] Now at Yahoo!
[3] Currently at the University of Michigan

0-7695-2507-5/06/$20.00 (C) 2006 IEEE

2. Background

As shown in our earlier work, sensemaking is the process a reader/analyst goes through when trying to collect, organize and represent the content of a document corpus. [13] As described in that earlier work, sensemaking is intrinsically iterative and creative as the analyst works through content, iterating both to restructure the available content (including or excluding content), as well as to create representations of the content for organizational and inferential reasons.

The sensemaking process is one that has several distinct activities—looking for instances of data, trying to create an organization of the data and using the collected data to resolve the questions at hand. Each sensemaking activity is characterized by a set of behaviors, each behavior having a cost (in time or energy) and expected utility as the analyst looks for additional information, considers alternative representations or organizations of the material, and evaluates the value of a document to the interpretation that is being constructed.

In many situations sensemaking characterizes a great deal of the work of analysts as they maintain background situation awareness and transition into tasks that answer specific analysis requests.

Our goal is to understand how people come to understand a collection of materials—in particular, what the effects of tools and methods are on the people who use them to organize materials and draw value from them. Here we focus on basic principles of organizing tools—contrasting information visualizations, paper piling, and automatic e-clustering tools. This work is not an attempt to find the single best tool or to provide an evaluation, but to find basic truths about how people organize their materials and the ways they interact with those materials.

3. Being literate with piles and clusters

While there have been a number of studies of the ways in which people create and use piles of personal information [6, 13], and a number of systems implemented to give piling capability to computer users [1, 2, 7, 18, 14], there has been remarkably little attention paid to the details of this behavior. People who pile, and the computer systems that support them, are simply assumed to be helpful as natural extensions of the way people natively work with document collections.

There are two notable exceptions: First, Barreau and Nardi's work [3, 9] considers ways people manage electronic documents at work. They identified three types: ephemeral, working and archived documents. Their study highlighted the overhead costs involved in managing individual hierarchies, such as laying out icons on the desktop and filing email messages into folders.

Second, Pirolli and Card [11] analyze the operations of a business intelligence office by examining document flows from pile to pile and throughout the office space. In this case, piles have fairly well-specified semantics and preserve their identity within an analytic framework over time.

A consistent finding among these studies is that action items associated with ongoing tasks are most readily at hand, often in stacks and piles on office surfaces. At the same time, personal archives are located within fairly quick access, and archival information is stored or available at further distances at greatest costs in time and effort. This suggests that even in commonplace settings, people are acutely aware of the costs of document access, and structure their environments to reduce the overall use costs.

3.1 Tradeoffs in working with collections

People who create piles or use document browsing tools are subject to many kinds of costs in their use. Even piled documents, which we tend to think of as an essentially unsophisticated operation, has an associated set of skilled techniques that are often used to organize and encode knowledge about their internal contents and structure. [5, 6]

Thus, even simple piles can be thought of as representational structures that are especially useful in intelligence analysis tasks where the structure of the document corpus may not be known, and the task can vary tremendously from day to day. What kind of costs do particular tools incur?

All representational structures have a set of things they do well to support a task, and contrariwise, a set of things that are difficult or nearly impossible to do. [19] For the kinds of sensemaking tasks we consider here, the relative cost structures of reading, browsing, finding instances of data, creating representational structures (even something as simple as an organizing pile) are all potentially significant costs in the overall performance of the analyst. [13]

3.2 Kinds of documents

There are tradeoffs that are made in the choices of analysis tools, methods and operations—although users tend to be less aware of the tradeoffs than they are of the day-to-day frustrations of use.

For instance, paper-based analysis methods have much to recommend them: paper documents allow for very fast and simple highlighting, annotation and pile-style organizational structures. Paper collections

can be easily manipulated, and in general, the use of paper collections reduces the overall level of distractions (few paper collections suddenly break with an illegal operation). Paper even affords a degree of multiple categorization, as a document can be placed midway between existing piles to indicate a measure of joint membership.

By contrast, electronic filing and clustering systems can handle extremely large quantities of documents, scaling far better into thousands and millions of documents without overtaking physical space and capacity. E-documents can be searched and clustered automatically, with explicit methods for multiple categorizations.

However, computer-based clustering methods are not cost free—they may be fast, but the end-user still needs to spend time understanding what the clustering system has done, evaluating whether or not it reveals a structure that makes sense, and if not, then the user begins a potentially long and complex negotiation with the system to correct the misalignment between human conception and e-clustering organizations.

Ultimately, our goal is to understand the basic characteristics of what helps analysts understand the content of a corpus, independent of the tools used. So we began a series of studies to get at the varying effects of tool use on the analysis process.

4. The Studies

In this section we relate three studies that we conducted to understand the detailed cost structure of people doing analysis tasks on large document corpora. We describe each study, then give a brief analysis of the results, followed by a rationale for the next investigation.

We began our studies with an assumption so basic that it seemed to scarcely merit investigation: that computer-based visualizations would help one understand a large document corpus. Yet when we ran this first experiment, the results were surprising. To put these results in context, we first remind the reader of this earlier work in Study #1 (reported in greater detail in [15]).

4.1 Study #1: Grokker1 Visualization

We believe that we need to understand sensemaking in realistic information understanding tasks. That is, we need to study tasks that are reflective of actual practice and not try to dissect out tiny individual subtasks such as just query formation or just reading comprehension. Towards this goal, we created a study to understand how people perform when faced with more documents than could be read in the allotted time; a situation that is uncomfortably common in many real-life situations (especially in many analytic tasks such as intelligence, strategic or business analysis). We call this kind of task the *grokking experiment* [15] as we measured the ability of a subject to *grok*, or deeply understand, a complex corpus. Performance was judged based on how much knowledge the subjects could internalize during the study period and was quantified by a written post-trial test. Subjects (N=12) were shown representative questions (for a city not used in the assessment test) before the trials, and did not know the specific questions during the trial.

Subjects were given collections of 300 documents, each a news article from one of 6 large international cities. The cities chosen were selected for their relative obscurity with respect to our test subject population in order to minimize the effects of background knowledge. We asked each subject to study the collection of documents for a short period of time. Then after either 5 or 15 minutes we tested their understanding performance. The times were much too short to allow careful reading—the subjects averaged only between 1 to 3 seconds per document. Subjects had to read, form a mental model, internalize the information, and then be able to answer questions after the trial period.

Subjects saw the collection of documents in one of three different forms: paper, semantic and temporal. In the paper form, each subject was presented with a bound collection of paper articles, numbering as many as 500 pages. (The paper document collection was bound, meaning that it could not be reorganized by the subject.) For the semantic and temporal displays we created a tool, *Grokker1*, that would show the data in one of two straight-forward visualizations. In the *semantic display* (see Figure 2), small rectangular icons were laid-out on the screen based on a simple 2-dimensional latent-semantic indexing (LSI) calculation. In the temporal display, the documents were arrayed by publication date in reading order—with the earliest article at the upper-left corner of the screen and the latest article at the bottom right. In both electronic presentations, when the mouse pointer was over one of the buttons, the first 100 words of the article were quickly displayed on the screen—a rapidly displaying tooltip that acted as a brief summary of the entire article. The user could click on the button to see the complete article in a new, separate, persistent window.

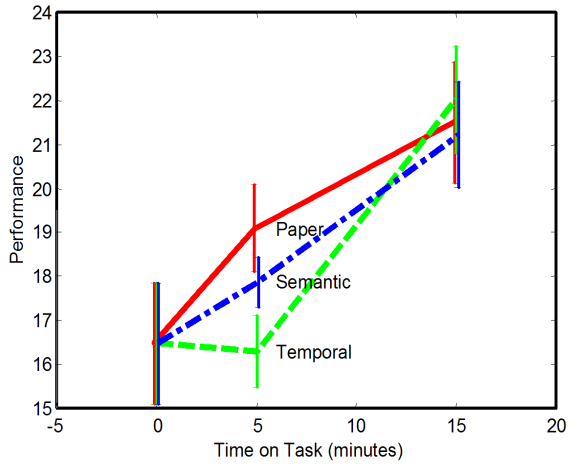

Figure 1. Sensemaking performance (questions answered correctly) using three different browsing techniques.

The results of the first grokking study are shown in Figure 1. All subjects, as expected, started with some baseline knowledge about each city. We measured their baseline performance with "time on task" at zero minutes; that is, we measured background information on each of these obscure topics. As one would expect, as they spent more time with the collection their scores, as measured by the number of questions they got right on a post-reading testing, went up.

We had anticipated that the electronic presentations would be significantly better than paper at aiding our subjects sensemaking. Both visualizations were designed to be simple, informative, and very fast and responsive. Both visualizations were intended to allow for fast skimming of the document collection at an abstract level (by brushing over elements in the display), and letting users drill down (by clicking on the elements) to show the details of the article.

Yet subjects appear to be much better at understanding the material using a large bound collection of paper. There are many possible reasons for this, including screen size and resolution, paper handling ability and overall familiarity with the paper medium. We were surprised at how the use of common visualizations did not seem to help much, and actually significantly decreased performance at shorter (i.e., more pressured) time intervals.

4.2 Study 2: Manual vs. computed

In a second study we observed how subjects (also members of the IBM research staff well-versed in document handling and research methods) performed a less structured task: reading, organizing, and preparing a presentation about a collection of documents.

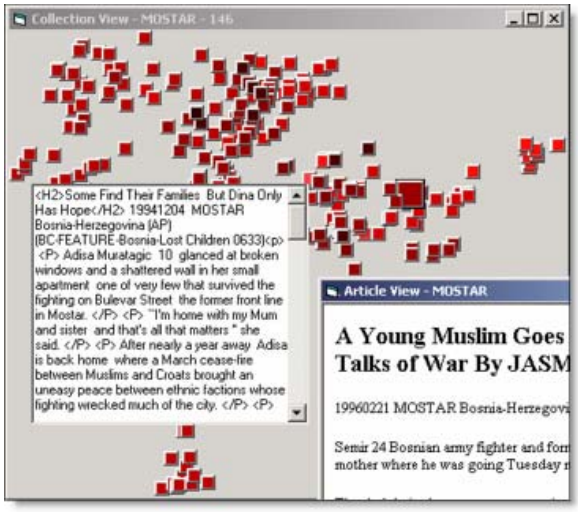

Figure 2. The Grokker1 semantic display shows articles as small square icons that are fixed in place as laid-out by latent semantic index dimensions. A small popup (on left) shows the effects of a roll-over, while a full article can be had by double-clicking on a document icon (right).

We were interested in how people browsed through the collection, organized the data, how they understood what was present, and what they wanted to read. We wanted to understand the differences between the paper and the electronic presentations. We video taped 10 subjects performing a task that was described to each subject as follows:

> Imagine you are an assistant of Jack, a senior analyst who gives advice on US' foreign policy in Asia. Jack is currently working on a case related to Azerbaijan. He has about 100 news articles in the period from 1994 to 1996. Jack wants you to find out what had happened in that two-year period. And he wants you to dig out the relationships among different events, different countries, etc. He has given you 1.5 hours to do the task. During the task, you are asked to write/draw down your findings and after the task, show them to Jack in a short 10 minute presentation. Please try to organize the findings in a way that is easy for others to understand the complex issues, especially the rich relationships among events, countries, etc. You may need to read some articles in detail in order to find subtle relationships. You are allowed to re-organize the articles in whatever way that helps your task. You are encouraged to bring out hypotheses on various relationships and show evidence to support them or show the rationale behind your hypotheses.

We studied subjects using paper documents and two different electronic systems: eClassifier [17]] and SSIGS [12]. (See Figure 3 for an image of SSIGS. For the purposes of this study, eClassifier is operationally similar.) eClassifier is a commercial product that automatically clusters documents based on their semantic contents and then displays the organized documents in a set of flat groups. SSIGS is a similar research tool that was designed to support the sensemaking task by providing a framework for organizing searches and their results.

In all cases (both paper and with both clustering tools) subjects were presented with 100 newswire articles about either the city news of Baku or the oil pipeline situation in Azerbaijan. Unlike our first study, the paper documents for this experiment were unbound, allowing readers the ability to create piles and clusters on the (physical) desktop freely. In all cases, subjects were pre-screened to ensure that the content was novel and unfamiliar, giving them very little background knowledge to influence their analysis process.

Each subject was then given 90 minutes to read through the collection, creating clusters in any way they wished (including not creating clusters, if desired) until they were satisfied that they understood the document corpus well-enough to give a five minute briefing to the mythical "Jack" of the problem statement.

The electronic systems had a number of capabilities that were not possible with paper. In this experiment, the eClassifier system initially clustered all 100 documents into 9 distinct clusters. Users were then free to move documents around within the clusters as they felt necessary.

With the SSIGS system, users could specify how many clusters they wanted: one of our users asked for 5 clusters and the other didn't use the clustering. In both cases, documents were organized on the screen and users could view short titles before deciding which articles to read in depth. In the statistics that follow we talk about skimming or hovering time as the time users spent reading the title of an article. (We judged this by asking users to point or talk about what they were reading.)

Evaluation: Initially we considered trying to evaluate the quality of the final output briefings, in practice nearly all the briefings were of high quality and indistinguishable. Instead, we focused on understanding the important differences in the way each of the tools were used and how that behavior affected the ability of users to read and gain insight into the collection.

The interaction techniques were quite different between the paper and the eClassifier or SSIGS tools. Figures 4 and 5 captures some of these differences: documents could be moved easily and rapidly into piles, document summaries could be browsed much more quickly with the tooltip rollover technique. Using the video of subject behavior created during each test, we analyzed the actions of our subjects, measuring how long different activities took, and how often they were performed.

Most striking is how much easier some activities are than others, and how this changed people's behavior. In the paper case, opening a document (picking it up from a pile of paper) is easier than with either electronic tool. This low cost / high ease of use is likely the reason that subjects read more articles on paper than they do electronically. This is true even when one adds in the number of articles where the subject just skims the titles. Subjects read more when they can access more of a document more quickly.

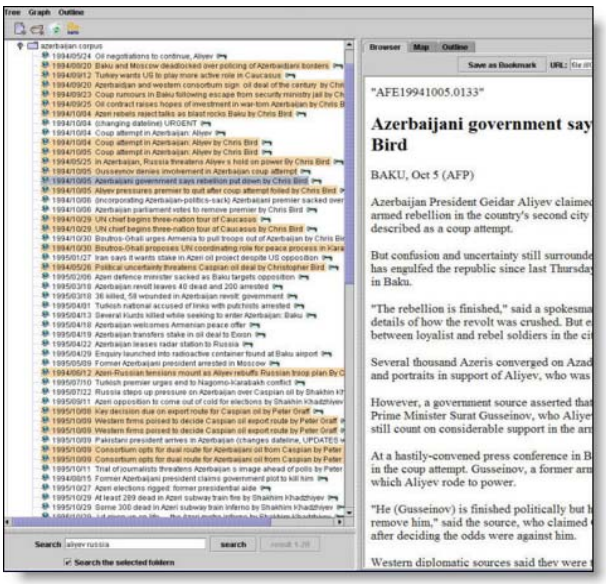

Figure 3. SSIGS automatically cluster documents into similar groups. An individual document can be read by clicking on its entry in the list view. It can be moved to another cluster by drag-and-drop into a different folder.

This behavior is consistent with the ideas in Gray [4] where subjects preferred using their imprecise memory of the task's target state to taking the fraction of a second to check the data that was available on their screen (but perhaps partially occluded). With some tools, moving a document is hard (in terms of time), so subjects found other ways to accomplish their sensemaking tasks. Perhaps keeping a mental model of where errant documents where, or adjusting their internal description of the cluster names to fit the organization provided by the electronic system.

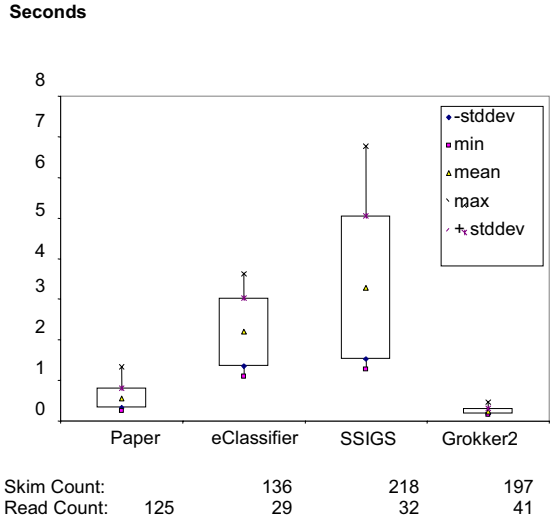

Figure 4. Time to read a new document in the paper setting, and to open a new document in electronic systems. After selecting a document, electronic tools display the full document to the user. The counts below the graph show the number of documents that are skimmed (by looking at the title summary in eClassifier and SSIGS, or the tooltip in Grokker2), and the number of documents where the full article is displayed.

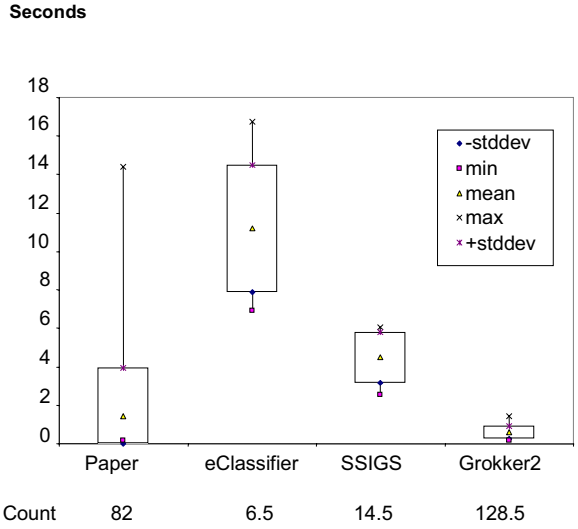

Figure 5. Time to move a new document and (below) the number of times a document was moved into a cluster in each of the four different conditions. The extreme variance of the paper case is caused by a single individual on one particularly problematic document.

4.3 Study #3: Redesign of Grokker1

In response to our subject's difficulties in Studies 1 and 2, we re-designed the original Grokker1 application to reduce the cost of operations for which paper was easier. Grokker2 facilitates both skimming and organizing the document collections. Figure 6 shows an image from the screen.

Grokker2 has four major changes to facilitate direct manipulation of the documents compared to the original tool: (1) Small iconic buttons have been replaced with larger buttons that contain the first few words from the article's title. This is enough to give the user a sense for the document's content without filling the entire screen. (2) Document icons are moveable. Users can quickly drag an icon anywhere on the screen: either to remove a document from consideration by moving it out of the way, or quickly sorting a document into a new pile. (3) Users can add text to the display to label the piles or organize their work. Finally, (4) we improved the formatting of the tooltip article display to make it easier for subjects to grasp the article's content at a glance. We limited the number of articles to 100, to match the collection size we used in Study 2.

As shown in Figures 4 and 5, the changes in behavior were dramatic. Subjects (N = 6) moved significantly many more documents with Grokker2 than they did even with paper. More interestingly, even though the time to access the full document was as fast as paper, if not faster, subjects were content to use the short summaries that were provided on average 197 times and only read the full article 41 times. The tooltip summaries were judged to be more useful by subjects since they were content to read the summaries, even though bringing up the full document was faster than paper. This was evidently because the tooltip summaries were even faster to access.

5. Analysis

From our detailed video analysis of the way people use physical piles to understand a collection, it became clear that people have very high facility for using paper news articles: the cost of access is very low, a person's ability to skim the article and get the gist is very fast (< 1 second), and paper can be sorted into informal clusters (piles) at a very high rate (~30/minute). Further, the creation of representations—even hierarchical clusters defined by positioning on the physical desktop and overlapping layouts—is quite good. In our experimental setup, a desktop of 1 meter x 0.9 meters could easily support 20 different clusters, with room for subtlety in the representations by positioning and twisting of the

piles. More importantly, since the clusters were self-defined (by direct manipulation with the user dragging document icons around while browsing through a collection), the analysts didn't require time for additional study to understand what was in the cluster. Informality worked to the user's advantage.

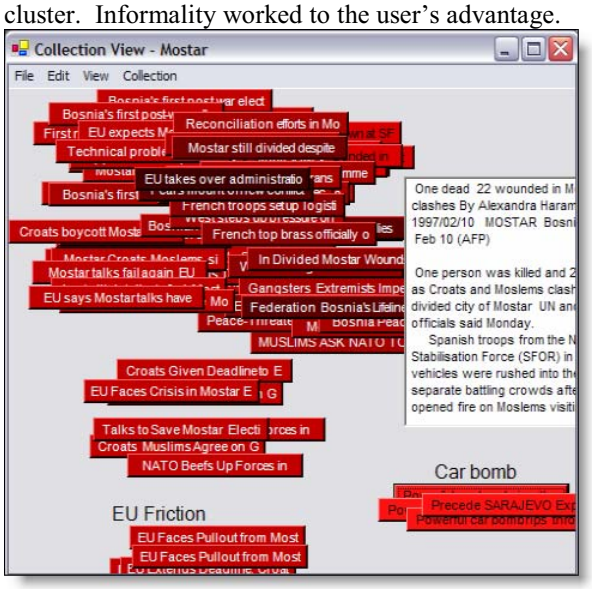

Figure 6. Grokker2 display created by a user. Each rectangular icon represents one document; the tooltip window is shown for one document as the user hovers over a document. This display has much more immediately available information for the reader to use.

By contrast, scanning is easy in the electronic tools but hard in paper. Grokker2 is successful at this task because by using direct manipulation, it reduces the time it takes to access and move a document, by as much as a factor of 10. This has a direct impact on the number documents that are moved between. With the re-design of Grokker (from Grokker1 to Grokker2) to more closely model the cost structures of paper, we found that users indeed act more as though they were working with the facility, grace and speed of paper.

In the physical world with paper, physical layout can be used to resolve the differences in ambiguous or as-yet undetermined categorization. This physical layout ambiguity is an important resource for representations, particularly when the representation is evolving or emerging from analysis. Adopting a interface design which is analogous to a physical desktop, Grokker2 enables such ambiguity as well as clusters with clear boundaries.

Subjects using eClassifier and SSIGS made a relatively small number of re-organizations (6.5 in eClassifier, 14.5 for SSIGS) after have the system automatically clustered the articles. Why was that? Were the clusters good enough, or was the cost of reorganization so high as to dissuade reorganization? Our experiment on Grokker2 answered this question by showing a surprisingly high number of re-organization movements when the subjects were given cluster structures on the interface. With the extreme low cost of moving in Grokker2, subjects freely reshape the existing semantic clusters to fit their needs.

A key insight in these studies is that the cost structure for using each of these tools and representations is very different. Figures 4 and 5 illustrate this point: the most common operations of accessing, moving and organizing documents vary tremendously from system to system—with the consequent effect that the number of documents seen, managed and used are very different, even in so small a study over such a short amount of time.

6. Conclusions

Clearly, the big disadvantage of physical documents is that the methods won't scale for extremely large document collections or very high flow rates. Obviously, automatic clustering and the use of tools to help organize large document collections will dominate any cost structure when the number of documents gets sufficiently large.

How can we take advantage of these studies? It became strikingly clear that the cost structure of paper documents and pile use has several distinct performance advantages: the time cost of directly accessing the contents of a document, the time cost of creating an informal (but highly useful) cluster, and the time cost of assessing an existing cluster. All of these actions are strikingly rapid, partly due to the physical affordances of paper, but also because of years of practice in reading newspaper articles on the part of our test subjects. We find that small changes in the time properties of these actions can cause dramatic effects in the ability of a user to see, manage and understand the corpus.

As Gray [4] has shown, even milliseconds matter when it comes to making tradeoffs between choosing to look for information available on the desktop or to access an internal memory. When faced with many thousands of milliseconds difference in the interface designs of our tools (e.g., the difference between eClassifier and Grokker2), we find huge differences in the number of documents seen and understood by the subjects.

Our plan is to continue to study these behavioral tradeoffs that are made by analysts on the basis of the interface properties and their effect on the cost structure of sensemaking. In the process, we hope to identify additional behaviors and interface designs

that will be able to significantly amplify the analyst's ability to work with and understand extremely large document collections.

7. References

[1] Ashdown, M., Robinson. P. 2005. The Escritoire: A Personal Projected Display for Interacting with Documents. *IEEE Multimedia Magazine*, 12(1): 34—42.

[2] Bauer, D., Fastrez, P., Hollan, J. 2005. Spatial Tools for Managing Personal Information Collections, *Proc. Hawai'i Conf. on Systems Sciences (HICSS)*, Kona, HI, 104.

[3] Barreau, D., Nardi, B.A. 1995. Finding and Reminding: File Organization from the Desktop. *SIGCHI Bulletin*, 27(3): 39-43.

[4] Gray, W. Fu, W-T. 2004. Soft Constraints in Interactive Behavior: The Case of Ignoring Perfect Knowledge In-The-World for Imperfect Knowledge In-The-Head. *Cognitive Science* 28, 359—382.

[5] Kirsh, D. 1995. The Intelligent Use of Space. *Artificial Intelligence*. 73: 31-68.

[6] Malone, T. 1983. How do People Organize their Desks? Implications for the Design of Office Information Systems. *ACM Transactions on Office Information Systems* 1(1) 99-112.

[7] Mander, R., Saloman, G., Wong, Y.Y. 1992. A 'Pile' Metaphor for Supporting Casual Organization Information. *Proc. ACM CHI '92*, pages 627—634, Monterey, California, May 1992.

[8] Marshall, C. S., Shipman, F. M. 1997. Spatial Hypertext and the Practice of Information Triage. *Proc. Hypertext '97*, 124-133, Southampton, UK, April 1997.

[9] Nardi, B., Anderson, K. and Erickson, T. 1995. Filing and Finding Computer Files. *Proceeding of. East-West Conference on Human- Computer Interaction*, Moscow, Russia, July 1995, 4-8.

[10] Olson, J., Olson, D. 1990. The Growth of Cognitive Modeling in Human-Computer Interaction since GOMS. *Human Computer Interaction*, 5(1) 221-265.

[11] Pirolli, P., Card. S., 1999. Information Foraging. *Psychological Review*, 106: 643–675.

[12] Qu, Y. 2003. Sensemaking-Supporting Information Gathering System. *ACM CHI'2003, Extended Abstracts*, 906-907.

[13] Russell, D. M., Stefik, M. J., Pirolli, P., Card, S. K. 1993. The Cost Structure of Sensemaking. *Proc. InterCHI '93*, Amsterdam, Netherlands, 1993, 269-276.

[14] Shen, C. Lesh, N. Vernier, F., 2003 Personal Digital Historian: Story Sharing Around the Table, *ACM Interactions*, 10 (2) 15-22, March/April.

[15] Slaney, M., Russell, D., 2005. Measuring Information Understanding in Large Document Collections. *Proc. Hawai'i Conf. on Systems Sciences (HICSS)*, Kona, HI, 105.

[16] Soper, M. E., 1976. Characteristics and Use of Personal Collections. *Library Quarterly*, 46, 397-415.

[17] Spangler, S. Kreulen, J. Interactive Methods for Taxonomy Editing and Validation. 2002. P*roceedings of the eleventh international conference on Information and Knowledge Management, CIKM 2002*, McLean, VA, November, 2002, 665-668.

[18] Wellner, P.D. 1993. Interacting with Paper on the DigitalDesk, *Communications of the ACM* 36(7) 87—97.

[19] Zhang, J. 1997. The Nature of External Representations in Problem Solving. *Cognitive Science* 21(2), 179—217.