

Measuring Information Understanding in Large Document Collections

Malcolm Slaney and Daniel M Russell
IBM Almaden Research Center
malcolm@ieee.org, daniel2@us.ibm.com

Abstract

We present a method for testing subject's performance in a realistic (end-to-end) information understanding task—rapid understanding of large document collections—and discuss lessons learned from our attempts to measure representative information-understanding tools and behaviors. To further our understanding of this task, we need to move beyond overly constrained and artificial measurements of easily instrumented behavior. From observations, we know information analysis is often performed under time pressure and requiring use of large document collections. Instrumenting people in their workplace is often untenable, yet oversimple laboratory studies often miss explanatory richness. We argue that studies of information analysts need to be done on tests that are closely aligned with their natural tasks. 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.

1. Understanding Information in large collections

How do people understand our complex, information-rich world? Here, we look at an aspect of the information-understanding problem, specifically, how well different tools help someone read, 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: creating the collection, scanning it, reading specific items and forging an understanding. This paper describes a test for measuring information understanding in an open-ended, semi-directed task, and describes some initial results.

Information understanding is a common and demanding problem. We are all information analysts and need to understand new information—for example, what causes knee pain? Quite often in the professional world, analysts 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 information understanding is

often done under extreme time and performance pressure [16].

Information understanding includes many well-studied problems [9, 10]. Information understanding, as we think of it, includes these five factors: information retrieval, reading speed, information foraging, sensemaking and question answering.

Information retrieval (IR) is a well-studied area, with well understood means of performance assessment. The goal of IR is information understanding, yet the traditional performance-measuring approach is to evaluate the precision and recall of the results from a single search [19]. Clustering and relevance sorting are often used to further organize the results. Information understanding 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. IR today is almost never done with a single query [21] and any number of documents can lead the user to a new understanding.

Rapid **reading**, or skimming, is a fundamental aspect of our experiments. Reading speed on a proofreading task is slower on old CRT screens compared to paper [7]. But a more recent test that replicated the electronic experience by having subjects read screen-sized chunks of text showed identical performance for paper and moderate-resolution screens [6]. Subject's reading speed varies over a wide range, more so when non-native readers of English are considered. We want to know whether the affordances of an electronic display will overcome the inherent advantages of paper.

In the **information foraging** analytical framework, information analysts search for information and decide how much time to put into one document, before deciding to move on, and selecting a new source of information [15, 17]. 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.

Sensemaking is the process of a person coming to understand a large body of information [18]. Each portion of the sensemaking process comes with a real cost. In sen-

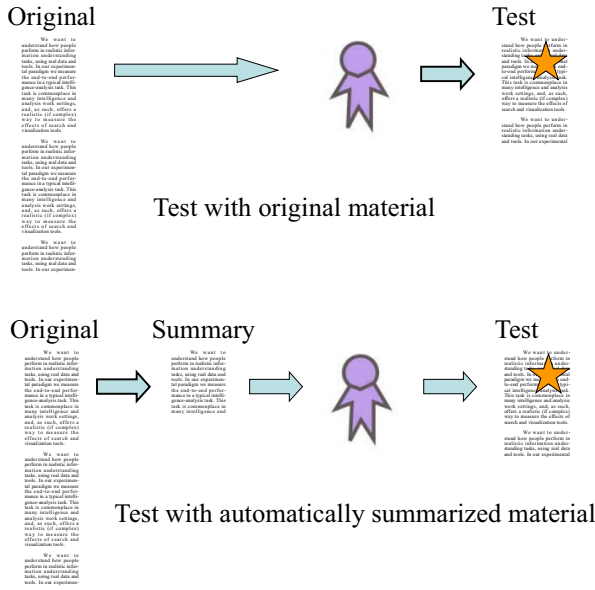


Figure 1:A schematic of a common machine-summarization testing procedure. A subject should get the same score on the test whether they read the original or the summary text.

semaking, 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.

Question answering (e.g., [23]) is one aspect of information understanding. Yet this paradigm is limited since analysts often 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. Our tasks are defined in terms of how well a human subject understands the material.

This work steps back from these individual tasks and looks at the bigger picture. How well can human subjects search, read, 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.

In many ways, an analyst’s task is similar to a document-summarization system. A system (or an analyst) must understand the gist of a document, so it (or the person) can write a summary. Document-summarization systems are tested in two ways: 1) a judge reads the summary and gives the summarization a subjective score, or 2) a tester reads either the original document or the summary and then is tested on the material. Ideally, testers get the same score whether they read the original or the summary (See Figure 1). We are not interested in the quality of the sum-

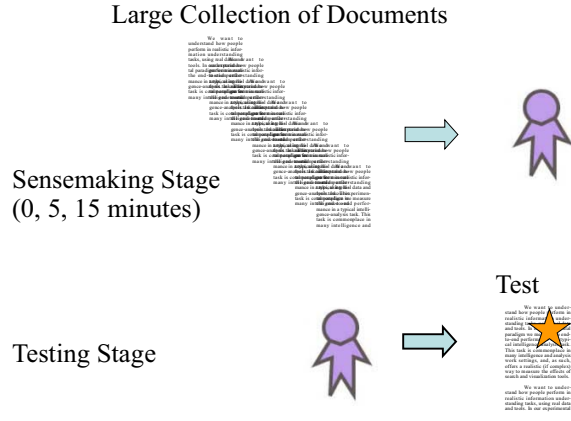


Figure 2:In our study, subjects first study a large document collection, creating their internal representation of the material. Then the documents are put away and the subject’s understanding is tested.

mary prose, so we ask our testers to form their own summary, and then we test this mental model. (See Figure 2)

The primary contribution of this paper is to describe a complete test of human information understanding capabilities. As Marti Hearst states [9, p. 262]

Empirical data involving human users is time consuming to gather and difficult to draw conclusions from. This is due in part to variation in users’ characteristics and motivations, and in part to the broad scope of information access activities.

We agree and argue that the end-to-end information-understanding test is important. We address user differences by characterizing the range of abilities. We address user’s motivations by providing rewards and a common information understanding goal. With a well-controlled test the human effort is manageable. The end-to-end test, in combination with the tests for the specific behaviors listed above, provides an important tool as we design and evaluate information-understanding efforts.

2. Creating a test to measure information understanding

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

Our test measures how well subjects understand a large number of documents when given different amounts of time. 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 and describe our results.

As is well known, **understanding** (or **comprehension**) encompasses a wide number of different phenomena [12]. The depth of understanding varies from person to person, as reader self-expectations vary tremendously, and from time to time, as task goals, interest and attention vary. In this paper, “understanding” a large document collection means that a reader has a general sense of the topics covered in the texts, has read some of them in detail, and is both able to answer factual questions and evaluative questions about the concepts within the texts.

2.1 ULDC—Understanding large document collections

The ULDC test we describe here measures the ability of subjects to understand a large document collection after spending varying amounts of time with the collection (where they can skim, scan, read, take notes, etc.), and then are given a multiple-choice test to measure their level of 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 on an 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 the ability to understand the documents in a news collection.

We tested three simple presentation tools for organizing, accessing and exploring a collection of documents: (1) a bound paper printout of the documents, (2) a temporal display and (3) a semantically-clustered overview display.

The essence of our experiment is as follows: We want to measure how much information a subject can absorb about a topic from a collection of documents using a particular information presentation. It is difficult to ascertain how much a subject knows about a topic without biasing browsing behavior. Thus each topic in our study consisted of news articles from international cities so we could assume that each subject knew relatively little about the city before the experiment.

Subjects create models of each collection in their heads, sometimes taking notes of their choosing, writing down important information on a pad of paper, as they use the specific presentation (or read the paper). We measure the subject’s success at the information understanding problem by testing the quality of their mental model through a questionnaire given them after a session of either 0, 5 or 15 minutes duration. (We measure separate subject’s background knowledge by giving them a test, with 0 minutes to read the document collection.)

In each trial of our experiment, a subject is randomly assigned a collection of documents and a presentation. Then after a set period of time the subject puts away the

Katmandu (Nepal)—Mountaineering, assignation, election, politics
 Bilbao (Spain)—Hostages, Basque, elections, explosions
 Baku (Azerbaijan)—oil, elections, visitors
 Vladivostok (Russia)—Military trial, visitors, Solzhenitzyn, general
 Paramariba (Suriname)—Drugs, logging, elections, general
 Riga (Latvia)—Latvian citizenship, divisiveness, general

Figure 3: Cities used in the information understanding experiments.

documents and is given a set of questions to answer. Their score is a measure of how effective that presentation is in facilitating understanding. We expect different presentations to give different performance, and for performance to increase as users spend more time with the collection. Because of learning effects, each collection of documents can only be used once per subject.

2.2 Data

We created six 300-article corpora and a set of questions for each. These corpora are large enough that subjects cannot read them in one sitting, but small enough to reasonably have an expert read them and write a test.

We used a selection of newspaper articles from the Giga-Word corpus collected by the Linguistic Data Consortium [13]. This corpus contains 4 million articles from four different international newswires. The newspaper articles are all professionally edited, and chosen by the original editors for their newsworthiness to an international audience.

Articles were sorted into sets based on their dateline city. Then we manually searched through the cities with between 300 and 2000 articles looking for appropriate candidates. We were looking for cities that had a handful of subjects. We didn’t want to use cities that were all about one subject (e.g. Papeete, mostly about Greenpeace protests over nuclear tests) or were a collection of random subjects (most of the articles from cities in China were government-issued stories on every possible subject.) We also skipped cities that had a large fraction of sports stories.

The articles and typical topics that we used came from the six cities shown in Figure 3. For each city, we then chose the 300 longest articles. These articles, in one of three different presentations, became the core document collections for our experiment. On average, each collection contains 150,000 words, much more than any subject could read in 15 minutes. This is small by IR or web standards, but is clearly a large problem for human subjects.

An outside contractor with an English-literature degree and many years of experience as an editor was hired to read a paper copy of each collection of articles. He then generated questions that would measure subject’s ability to perform the following task:

Your boss has just moved to town. What is important for her to know to be a knowledgeable member of the community?

5. During the mid-90s, roughly what percentage of Mostar residents left the war-torn city?

- a. 10-15 percent
- b. 45-50 percent
- c. 75-80 percent
- d. 90-95 percent

6. In 1995, French Legionnaires in Mostar were nearly killed from:

- a. Flash flooding
- b. Hurricane
- c. Train derailment
- d. Poisoned rations

7. In June 1996, a notable event for post-Yugoslav War Bosnia:

- a. First municipal elections
- b. Launch of first Bosnian airline
- c. Death of Mostar's acting mayor
- d. World Bank membership

Figure 4: Sample questions and answers for the city of Mostar. The correct answers are marked with an underline. (Mostar was not used in our tests.)

Several sample questions (and potential answers) are shown in Figure 4. Our question-generating editor spent on average four hours with each collection generating between 30 and 38 questions (and answers). We gave him enough time with each collection so that he did not feel the need to use the computerized displays.

The editor wrote questions that varied in style, but which tested factual understanding (as in question 6 of Figure 4), the ability to recognize slight variations on concepts presented, and the ability to provide a simple combining evaluation of multiple concepts over multiple articles (as in question 7 of Figure 4, where no single piece of text identifies the municipal elections as the most notable event of 1996 in Bosnia).¹

2.3 Presentations

The presentations we used in this study were designed to be representative, yet easy to understand and replicate. Many sophisticated and rich interfaces are described in the literature [9] and are available as commercial products. We wanted to understand the factors that lead to a good tool and didn't want the present study to turn into a bake-off of implementation details, or of the skill of an individual visualization user. These are factors we attempt to exclude from the study.

We looked at three simple presentations for browsing the collection: paper book, temporal visualization, and a semantic visualization. In each case, the presentation was as simple as possible, conveying just the basic information, but a realistic example from the literature. Clearly a real interface will combine elements of each interface with

1. We can provide the story identification tags from the LDC database and the tests we used to interested parties.

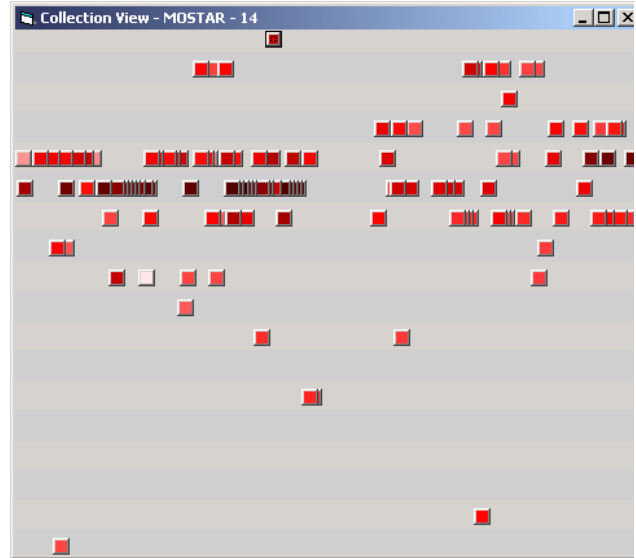


Figure 5: The temporal information display. Each article is represented by a single square in a linear temporal display, reading left-to-right across multiple lines.

easy-to-use filtering and search capabilities. Our goal was not to design the ultimate interface, but to understand the factors that lead to a good test.

2.3.1 Paper

The simplest presentation is a collection of paper. We printed each set of 300 articles on double-sided 8 1/2" x 11" paper and bound them into a book. The articles were sorted by newswire (AP, New York Times, etc.) and then by date. Each article was formatted with a bold headline, and then normal text for the article. There was no other formatting. Each article started at the top of a new page. A city's collection consisted of 350 to 580 pages of information. The advantage of the paper presentation is that there is a large amount of information, literally at the reader's fingertips. It is easy to skim and browse. The disadvantage of this approach is that it is difficult to get a sense of the big picture.

2.3.2 Temporal

The second presentation is a temporal display of the articles [22]. This presentation is shown in Figure 5. Each article is represented as a square, each of which responds to a mouse click, and represents a document. The articles are arranged in a single time line that is folded into multiple rows in reading order. Alternating light-gray bars were used to help subjects visualize the lines, and to visually break up any clusters across rows. Each article was placed linearly along the time line, and no effort was made to spread apart articles that were published on the same day. The color of the square is set along a dark-red, red, light-red scale and corresponds to the second LSI dimension (latent semantic indexing, see below). There is no date information in the display, save for the dateline embedded

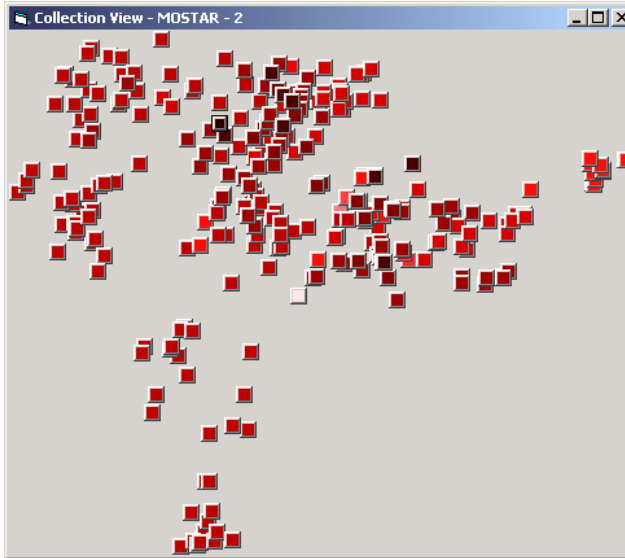


Figure 6: The semantic information display. Each article is represented by a single square in a spatial display that groups similar articles into clusters.

in each article, yet groups of articles often bunch up around certain dates (since they are news articles) and the temporal display does shows dates without news as large blank areas.

The advantage of the temporal approach is that many news articles on a similar topic are published near the same date so temporal adjacency suggests semantic clusters. The disadvantage of this approach is that many overlapping articles are hard to read.

2.3.3 Semantic

The third presentation is a semantic visualization of the articles similar to work described by others [e.g. 5, 24, 25]. This presentation is shown in Figure 6. Again, each article is represented as a square, but now the position of the square is based on a simple LSI analysis [2]. As in conventional LSI processing, we removed stop words (using the START list) and used a Porter stemmer before collecting word histograms of each document. The term-frequency matrix is adjusted using entropy global weighting [4] before an singular-value decomposition (SVD, using MATLAB) was computed. Conventionally, the singular values of the SVD are arranged in decreasing order, so the SVD approximation using the largest singular values is the best possible approximation to the original cloud of data.

The first SVD dimension is ignored since it largely corresponds to the mean of the data. Thus the horizontal axis in the visualization is mapped to the second SVD dimension, and the vertical axis is mapped to the third SVD dimension. The color of the square is set along the same dark-red, red, light-red axis and now corresponds to the date of the article.

Our visualizations are based on common themes in the literature. The semantic scatter-plot is similar to the Gal-

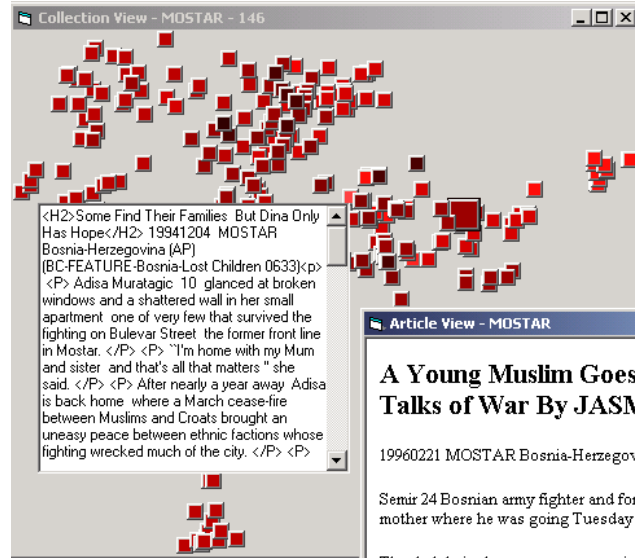


Figure 7: The semantic information display with a popup window caused by rolling over of an item with the mouse (left), and a second window showing the formatted article (right).

axy displays of Wise [24], StarField displays [11], and the Bead system [1]—these kinds of visualizations use a scatter-plot technique to represent documents as points in 2D or 3D layout that compresses a number of dimensions down into a tractable display.

A Microsoft Visual Basic tool was designed to present these visualizations. Each article is represented by a small square in a large window filling the screen. Two additional windows help the user to see the article in more depth. As the mouse moves over an article's square, a small window quickly appears showing a portion of the unformatted article. Only one of these popup windows is visible at a time (see Figure 7). If the user wishes to see more he can click on the square to see the entire formatted article in new window. The temporal and semantic displays were fixed during the experiment.

3. Methodology

We used twelve subjects, each browsing articles from six cities using one of three presentations, in our study. Each subject saw each city only once and saw all three presentations with both 5 minute and 15 minutes to study the city collections. These times were determined ahead of time based on a small pilot to give interesting differences. Both times were significantly shorter than the 4 hours used by the question-writer. Once a subject has studied a city, we can not use that city for that subject again since they no longer are starting with baseline knowledge.

The order of cities and presentations were randomized, counter-balanced and insured that all combinations of city and presentation-time were tested an equal number of times. Subjects were all post-graduate researchers in our

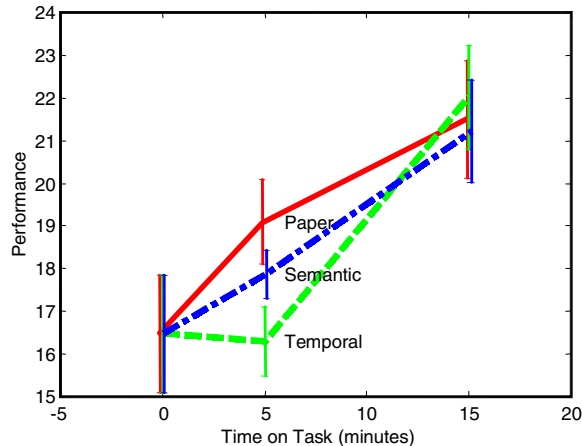


Figure 8: Performance of the subjects in our multi-document summarization experiment. Zero minutes represents our baseline data (when subject's did not access the articles.)

lab, all in computer science, engineering, or the physical sciences. They were compensated with gift certificates (\$10). To encourage our colleagues to take this test seriously, we gave the highest-scoring individuals up to \$15 extra. Subjects took from 2 hours to 2.5 hours to complete six trials. None of the subjects reported any special knowledge from any of the cities (one reported knowing somebody in one of the cities.) All but two finished all six trials at one sitting.

An additional twelve subjects took the same test without access to the documents, but based on what they already knew about the topic. This data represents our baseline performance measures, or zero minutes time-on-task.

4. Results

Figure 8 summarizes all the results of this study for the baseline study, and the 5-minute and 15-minute tests. Performance is shown as the number of correct answers per city test. Each data point represents the mean and standard error for 12 different subjects.

4.1 Time on Task

The overall trend of Figure 8 details 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 datapoints 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.

On average, performance increased from the baseline, to the 5-minute study, to the 15-minute study. With 5 minutes

of study, our subjects did not improve their scores using the temporal presentation, they were able to answer 1.5 more questions using the semantic presentation, and they answered 3.5 more questions using the paper collection of articles. After 15 minutes of study, our subjects using any of the presentations, could answer 4.9 more questions than the baseline subjects. Using a t-test assuming unequal variances, the 5-minute result ($M = 17.75$) is significantly better than the baseline result ($M = 16.5$) ($t(63) = 4.46$, $p < 0.001$). The 15-minute ($M = 21.6$) is significantly better than 5-minute result ($M = 17.75$) for all presentations ($t(84) = 1.96$, $p = 0.026$).

4.2 Visualization Tool

The results at 5 minutes for subjects using paper, semantic, and temporal tools are significantly different ($F(2,22) = 5.26$, $p = 0.014$). To our surprise, the paper presentation did best in 5-minute tests: paper is not significantly better than semantic ($F(1,11) = 1.905$, $p = 0.195$), paper is significantly better than temporal ($F(1,11) = 8.442$, $p = 0.014$), semantic is significantly better than temporal ($F(1,11) = 5.034$, $p = 0.046$). After 15 minutes, subjects performed identically with all three presentations

5. Discussion

The goal of our experimental design is to measure the degree to which a reader can work with a document collection and become generally more informed about the topic. As described earlier, this is a fairly common task in many situations. Yet in reviewing the literature, we were surprised that few task-critical measures of performance exist. Some authors (i.e. [14, 20]) suggest a narrow test. We argue that an end-to-end test is necessary to fully understand a system's behavior. Such a test will help other researchers better design and test their information retrieval and understanding tools.

The results of this paper can be summarized as "paper wins, but people persevere." Our initial belief was that the visualization tools used in this test would be so obviously better than the paper-based alternative, and so obviously simple in their use that we were simply waiting for the results of the t-test. We were surprised. Our results are consistent with those in the SuperBook trials [3] where computerized visualizations fared poorly in the least structured task (when questions did not contain words from the article or their titles.)

We had anticipated that the virtual visualizations would be significantly better than paper at aiding our subjects to understand the material. 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 the user gain drill-down information (by clicking on elements) to show the details of the article. With the visu-

alizations, multiple documents could be seen in detail simultaneously and we had hoped for more user insight into the topic areas by their investigation of the displayed semantic clustering structure.

Yet subjects appear to be much better at understanding the material using a large bound collection of paper. Some subjects complained that there was simply not enough information in the windows. It is worth noting that even with the large 21" monitor at the XGA resolution we used for our test, the printed page has a significantly higher information content. Our users were all highly skilled paper document users that could scan the bound paper documents quickly, recognizing salient features, and were able to use fingers placed into the bound document as temporary place-holders for quick comparisons

There is a chance that there is a bias in the test because the questions were prepared using the paper collection of articles. Our question writer felt that he knew the articles well enough from the bound volume that he did not feel a need to browse the collecting using other methods.

We attempted to use the baseline measurements and "subtract" the average subject's background knowledge, but our results became more muddy. In retrospect the reason is due to the noise that the baseline measurement adds to the calculation. The normal subject measurements and the baseline measurements each produce a noisy measurement. Since the subjects used in the two tests are independent, the variances of the noises add when the two numbers are subtracted. With many more subjects in the baseline experiment we can produce better estimates of the baseline knowledge, but then it is harder to ensure that the total population is uniform.

Although we recruited subjects from a fairly homogeneous researcher pool, the subject population was not as uniform as we hoped. Two subjects out of our 12 baseline subjects performed significantly better on the baseline tests than the others, and two subjects from the primary experiment performed significantly worse than the rest. While we speculate that raw reading speed might be an important determiner of final score, we cannot rule out population variations in international knowledge.

Subjects were given a written description of the presentations and a demonstration of their behavior. It was our initial belief that the visualization tools would be simple and straightforward enough to be obvious in use so that training and practice times could be minimized. But in retrospect, our testing setup probably did not provide sufficient practice time for subjects to develop effective strategies for use of these presentations. Subjects were allowed to take notes in whatever form they wanted to augment their performance, although this was not effectively used by any of the subjects tested.

Thus, we believe that *every* visualization needs testing—not just on toy tasks, but more generally in a realistic task setting that stresses the visualization and the ability of its human users to accomplish ecologically valid task goals.

6. Lessons learned

In the process of creating the ULDC test, we found that measuring information understanding 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 information understanding 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 hinder: measurements count, intuitions do not. The tools we gave our subjects (semantic, temporal and paper) were consciously designed to be straightforward instances of visualizations that are common in the literature. By not including many additional features in the toolset, we were hoping to measure just the intrinsic efficacy of the visualization method. As noted, we were surprised by how the subjects actually performed with these fairly generic tools.

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 population of information analysts. We need to accept the inherent variance in human background and abilities. We strongly recommend that all measurements of visualization tools (or of information understanding behavior in general) measure and characterize the effects of background knowledge and operator skill as part of the basic measurements of tool use.

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: A pilot experiment, with just six subjects, demonstrated the subtleties involved in the choice of test subjects. Subjects for the pilot experiment and for this visualization test were chosen from a population of full-time professional researchers at our laboratory site. 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 from the pilot 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 information understanding task, can be as much as an order-of-magnitude slower. The non-native readers of English were replaced by native readers of English in the study reported in this paper.

Textual material: The raw reading material we used for our study limits the kinds of comprehension questions we can ask our subjects. A study-guide for the GRE describes six kinds of comprehension questions: [8]

There are six types of reading comprehension questions. These types focus on (1) the main idea or primary purpose of the passage; (2) information explicitly stated in the passage; (3) information or ideas implied or suggested by the author; (4) possible applications of the author's ideas to other situations, including the identification of situations or processes analogous to those described in the passage; (5) the author's logic, reasoning, or persuasive techniques; and (6) the tone of the passage or the author's attitude as it is revealed in the language used.

In general, wire-service newspaper articles use a single focus, report explicit facts in a straightforward neutral fashion and do not persuade or argue. This study does not address all 6 types of comprehension. Yet our test measures more than fact retrieval—it also measures subject's ability to combine evidence from multiple documents.

Implications of the study: (1) It is important to measure consistent subject populations, rather than post hoc trying to 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. Information understanding 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) Information understanding tests must be conducted with a significant number of documents (to stress human limits, not just play to strengths of tool under study).

7. Summary

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

We have designed and validated an end-to-end test for information understanding tools. This test measures a subject's ability over a range of tasks that are necessary for information understanding. These tasks include: information retrieval and visualization, information foraging, sensemaking, and question answering. We ask subjects to build a mental model of the subject material by using a paper document collection, a scatter-plot visualization or a timeline (temporal) style visualization. The subject is then tested on the quality of their mental model by answering questions that require a deep, integrated understanding of the material. This is a realistic test for a common task that many information users have: the need to rapidly assess and understand a large document collection.

Our study demonstrated significant results as a function of time-on-task—starting before subjects used the tool (baseline) and measuring up to 15 minutes of tool usage. Most importantly, we saw significant differences between different presentation tools, demonstrating our test's ability to provide guidance as we design better tools. The tools in this study are simple—they isolate key components of information visualizations that are standard in the literature. Thus, we find that scatter plots that compress many semantic dimensions into a smaller space do not seem to provide any particular advantage for information understanding.

We believe that this style of test, which focuses on the key aspects of information visualization tools, used in a task that is rich and true-to-life, facilitates understanding of the effects tools have on their users, and provides an important baseline for future studies.

8. Acknowledgements

We appreciate the statistical help we received from Chris Campbell, editorial assistance from Lisa Yanguas and the IR insights from Marti Hearst. We are especially grateful to our very bright subjects, who endured a long study and clearly did not enjoy scoring so low on a test.

9. References

- [1] Chalmers, M., and Chitson, P. Bead: Explorations in Information Visualization. *Proceedings ACM SIGIR*, pp. 330–337, 1992.
- [2] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41 (6): 391–407, 1990.

- [3] Egan, D.E., Remde, J.R., Gomez, J.M., Landauer, T.K., Eberhardt, J. and Lochbaum, C.C. Formative Design-evaluation of SuperBook. *ACM Transactions on Information Systems*, 7 (1) pp. 30–57, 1989.
- [4] Dumais, Susan T. Enhancing performance in latent semantic indexing (LSI) retrieval. Technical Report Technical Memorandum, Bellcore, September 1990.
- [5] Feldman R., Kloesgen W. and Zilberstein A. Visualization Techniques to Explore Data Mining Results for Document Collections. *Proceedings of ACM SIGKDD*, pp. 16–23, 1997.
- [6] Garland, Kate J. and Noyes, Jan M. CRT Monitors: Do they interfere with learning? *Behaviour & Information Technology*, 23(1), pp. 43–52, 2004.
- [7] Gould, John D., Grischkowsky, Nancy. Doing the same work with hard copy and with cathode-ray tube (CRT) computer terminals. *Human Factors*, 26(3), pp. 323–337, 1984.
- [8] GRE Reading Comprehension Web Page. http://www.gre.org/practice_test/takerc.html. June 2004.
- [9] Hearst, Marti. User interfaces and visualization. In *Modern Information Retrieval*, Ricardo Baeza-Yates and Berthier Ribeiro-Neto, eds., Addison-Wesley, 1999.
- [10] Heuer, R. J. Jr. *Psychology of Intelligence Analysis*. United States Government Printing Office, November 1999.
- [11] Kandogan, Eser. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. *Proceedings of ACM SIGKDD*, pp. 107–116, 2001.
- [12] Kintsch, W. *Comprehension: A paradigm for cognition*. Cambridge University Press, New York, 1998.
- [13] Linguistic Data Consortium. The GigaWord Corpus. <http://www ldc.upenn.edu/Catalog>
- [14] Morse, Emile, Lewis, Michael, and Olsen, Kai A. Testing visual information retrieval methodologies case study: Comparative analysis of textual, icon, graphical, and “spring” displays. *J. Am. Soc. Inf. Sci. Technol.*, 53(1), pp. 28–40, 2002.
- [15] O’Day, V., and Jeffries, R. Orienteering in an information landscape: How information seekers get from here to there. *Proceedings of InterCHI ’93*, Amsterdam, Netherlands, 1993.
- [16] Patterson, E. Computer-supported inferential analysis under data overload. *Proc. CHI 1999*, Pittsburgh, PA, 2002.
- [17] Pirolli, P., and Card. S., Information foraging. *Psychological Review*, 106: 643–675, 1999
- [18] Russell, D. M., Stefik, M. J., Pirolli, P. and Card, S. K. The cost structure of sensemaking. *Proceedings of InterCHI ’93*, Amsterdam, Netherlands, 1993.
- [19] Salton, G., and McGill, M. *Introduction to Modern Information Retrieval*, New York, McGraw Hill, 1983
- [20] Sebrechts, M., Vasilakis, J., Miller, M., Cugini, J., and Laskowski, S. Visualization of search results: A comparative evaluation of text, 2d and 3d interfaces. In *Proc. of ACM SIGIR*, 1999.
- [21] Spink, A., Jansen, J. and Ozmultu H. Information seeking and mediated searching study. Part 3: Successive searching. *Am. Soc. Inf. Sci. Technology*, 53 (9), pp. 716–727, 2002.
- [22] Swan, R. and Allan, J., Automatic generation of overview timelines. *Proc. ACM SIGIR*, 2000.
- [23] Voorhees, E. M. Overview of the TREC 2002 question answering track. *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*, 2003.
- [24] Wise, James A., Thomas, James J., Pennock, Lantrip, David, Pottier, Marc, Schur, Anne and Crow, Vern. Visualizing the Non-Visual: Spatial Analysis & Interaction with Information from Text Documents, in *Proceedings of IEEE ’95 Information Visualization*, pp. 51–58, Atlanta, GA, Oct. 1995.
- [25] Wise, J. A. The ecological approach to text visualization. *Journal of the American Society for Information Science* 50 (13): 1224–1233, 1999.