In Depth Observational Studies of Professional Intelligence Analysts

Jean Scholtz and Emile Morse
National Institute of Standards and Technology
Gaithersburg, MD
{jean.scholtz; emile.morse}@nist.gov

Tom Hewett
Drexel University
Philadelphia, PA
hewett@drexel.edu

Abstract
Our goal is to produce metrics for measuring the effectiveness of software tools and environments produced for the intelligence community. To this end we need to understand the analytic process and to determine which data need to be captured to meaningfully measure process and effectiveness. In this paper we compare data from observational studies of professional intelligence analysts with data collected from an instrumented environment. We discuss some findings and their implications for possible metrics and for additional data needed to compute potential measures.

Introduction

The Advanced Research and Development Agency's (ARDA) Novel Intelligence from Massive Data (NIMD) program is undertaking programmatic research to design an environment to assist intelligence analysts in their work [1]. An important step in producing such an environment is to understand the process and products in the intelligence community coupled with an understanding of the cognitive demands on intelligence analysts. Today's analysts are faced with massive amounts of data and quickly changing worlds. They must be proactive in recognizing potential new threats to the safety and security of citizens.

It has been suggested that the intelligence community currently spends more time doing data collection and report generation than analysis and that the "bathtub" curve should be inverted [8]. This implies one measure of effectiveness would be lessening of the time spent for these two activities, leaving more time for analysis. Similarly, Heuer [6] points out that "more and better information" is not the answer to improving the analytic process and its results. Instead he proposes that greater attention should be given to more analysis of the information that analysts already have. For example, he suggests analysis of competing hypotheses as a methodology for organizing and analyzing information.

Our immediate concern is evaluation of a support environment for analysts. To properly evaluate the effects of software for analysts we must first devise appropriate metrics and then collect the necessary data to produce measures for calculating these metrics. In this paper we discuss an observational study of two professional intelligence analysts and compare the observations to data captured automatically by an instrumented environment. We discuss measures that can be calculated from the data. Additionally, we discuss some deficiencies in the measurement process and propose some possible solutions for additional data collection.

The Study

Two highly experienced intelligence analysts are employed by the NIMD program. These analysts are assigned tasks that can be completed using open source information. The analysts routinely work in an instrumented environment that captures the following time stamped information; 1) keystroke data; 2) search terms; 3) URLs visited; and 4) applications used. In addition, analysts may add notes to describe their tasks and may include comments directed towards aiding the researchers in understanding the data being collected. All data are stored in a relational database and may be accessed using database queries. In addition, the computer screen is recorded digitally and can be replayed to review the actions of the analysts. For the observational study we also recorded a video of the analysts' movements to catch offline activities such as reading hard copies of documents, talking on the telephone, times when they left the office, etc.

Our purpose in conducting the observational studies was to determine how much of the overall analytic process was being automatically captured and to determine metrics that would appropriately describe the analytic process. Observations were conducted over 2 consecutive working days. Two of the authors observed the two analysts, each starting with one analyst for a day and then switching to the other on the second day. Although this switch required the observers to quickly get up to speed on a new problem on the second day, we felt switching gave us a better overall view of the analysts. It also created a situation in which each analyst was indirectly encouraged to engage in thinking-out-loud about their strategy and progress for the new observer. Each analyst was given a different realistic short-term task at the beginning of the first observation day and asked to have a short report delivered to the observers by 3 pm of the second day, just as they are often expected to do when dealing with requests from their clients. This enabled the observers to see all phases of the intelligence cycle: tasking, data collection, analysis, and report generation [3].

On day 1 we met with the analysts to get acquainted and discuss procedure. We stressed that we were there to evaluate the automated data capture and that observations were to be as non-intrusive as possible. We sat in a corner of the analysts' offices positioned so that we could tell which application(s) the analysts were using (i.e., a browser, a search engine) but not close enough to read what an analyst was typing. We made "time stamped" notes when an activity occurred, but did not interrupt an analyst at work. Observer questions about activities were noted and "time stamped" so that in our end of the day debriefing session, we could review the screen recording or video tape and ask the analyst to retrospectively explain what had been happening at that specific point in time. With only about 1 min. granularity in our notes a number of activities show the same time stamp but can still be ordered in time. Figure 1 shows a small portion of the spreadsheet record of our observations. In our post analysis, we abstracted lower level behaviors into larger chunks, such as note taking, or phone call, as in the example below.

2003-10-09 10:49:00	1.00	offline	offline: looks at notes on pad	
2003-10-09 10:50:06	1.10			Note Making
2003-10-09 10:49:00	0.00	offline	offline: writes notes on pad	
2003-10-09 10:50:00	1.00	offline	offline: mentions that he wants to see if there is a connection between ...	
2003-10-09 10:51:00	1.00	offline	offline: makes notes on pad	
2003-10-09 10:52:00	1.00	offline	re-reads tasking	
2003-10-09 10:54:00	2.00	offline	makes notes on pad	
2003-10-09 10:59:00	4.00	offline	manual looked at	
2003-10-09 10:56:17	0.75			
2003-10-09 11:08:00	8.00	offline	break	Break
2003-10-09 11:06:53	5.54			
2003-10-09 11:09:00	1.00	offline	checks time - says will make call to CNS	
2003-10-09 11:10:55	1.93			Phone Call
2003-10-09 11:10:00	1.00	offline	log book to find phone number	
2003-10-09 11:11:00	1.00	offline	makes phone call- gets voice mail and asks to be called back	
2003-10-09 11:14:00	1.00	offline	makes a note on post it - pink:)	
2003-10-09 11:33:00	2.00	offline	explains about open source data and issues	

Figure 1: Observational Data Example

Results

After the observations were completed, the author who had not participated in the observations led a team in analyzing the automatically captured data. The goal was to keep initial analysis of the 2 streams of data, instrumented and observed, separate so that the eventual comparison would not be biased towards one or the other. We agreed in advance upon a number of measures that we wanted to calculate from both streams of data including; 1) time spent in the various applications; 2) time attributed to the various portions of the analytic process; 3) Gaps in data –off-line activities; and 4) report generation growth, The observational spreadsheets were separately analyzed using the same measures so that we could make comparisons and determine which data were being adequately captured and which data were being missed. We also want to determine what inferences we could make from the captured data to higher level abstractions of behaviors.

Observational Results

The analysts are referred to as analyst 3 and analyst 4 in the automated data capture environment so we use the same terminology here. Individual differences between the two analysts were quite apparent. Analyst 3 preferred to print out documents he was interested in reading while analyst 4 did most of his reading online. Analyst 4 copied and pasted from the online documents into the document that became his report in the end. Analyst 3 also did some of this but many of his notes were made on the hard copy documents and he later transferred these by hand into the final report. Analyst 3 worked alone for the entire task. Analyst 4, however, called in some additional expertise. He contacted an acquaintance by telephone and asked her if there was additional information to share that was not posted on the web site of her institution. He also assigned two junior analysts in the office to research a small, specialized portion of the task and had a face-to-face meeting, a phone conversation, and exchanged several e-mails with them. Their report became an appendix to his final product.

The most striking commonality in the behavior of the analysts was that on-going analysis occurred throughout the entire two working days. Both analysts frequently made comments about new information they had discovered and quite obviously used this information to guide future searches. We noted several "critical incidents" during the process. For example, analyst 3 had a breakthrough when he found a document that basically cut his task in half. He had been searching for information about potential activities of 6 individuals. After several hours he found a document that gave him reason to believe that 3 of these individuals had already been captured. Finding this information, however, was not straightforward as he had to have knowledge of aliases and alternative spellings of the names.

Quantitative Results

We have both days of electronically captured data from analyst 4, but due to technical difficulties only the 2nd day's data is available from analyst 3. Table 1 shows the time comparisons for the observational data and the electronically captured data. The offline time shown in Table 1 is actually inflated from working off-line time as it included time both for morning coffee and during the lunch hour. (During these time periods the analysts and observers did not discuss the particular task. Rather, any discussion was either social or confined to discussion of study procedures and general analytic strategies.) Also, the differences in total time are due to the fact that the analysts sometimes logged into the instrumented environment prior to the start of observations and/or logged off after we had finished the afternoon debriefing session.

	Observational data				Instrumented data			
	Day 1		Day 2		Day 1		Day 2	
	Total time	Time offline	Total time	Time offline	Total time	Time offline	Total time	Time offline
Analyst 3	5.77 hours	3.88 hours 67%	5.02 hours	3.22 hours 64%			6.69 hours	5.00 hours 75%
Analyst 4	6.05 hours	3.1 hours 51%	7.95 hours	4.12 hours 51.78%	7.68 hours	4.38 hours 57%	7.32 hours	4.52 hours 62%

Table 1: Comparison of Observation time/ Instrumented time

The comparison of most interest in Table 1 is the time offline. For analyst 3 on day 2, the observed offline time was 3.22 hours while the automatic collection logged 5.00 hours. For analyst 4, the observed time on day one was 1.2 hours less than the automatically logged time; on day 2 it was 0.4 hours less.

Figure 2 shows the graph for time gaps for analyst 4 on day 1. Note that some of the deviation comes from our definition of "gap" in the automatic collection data. We defined a gap as 30 seconds or longer with no keyboard activity. Nevertheless, by looking at the graph, it is reasonable to infer that the analyst is not at his desk during the very long periods with no activity. What activities occur in the shorter gaps? Due to human subject regulations, the e-mail activity and phone conversations of the analysts are not captured electronically. Similarly the face-to-face meeting that analyst 4 had with the junior analysts was observed but not captured electronically. The time when analyst 3 went to the printer to collect documents was not noted as such nor was the time he spent reading off-line. However, to some extent these activities could be inferred. For example, a period of non-activity directly after a print command was issued. Finally, some activity was not recorded either by the observers or by electronic capture. For instance, Analyst 4 arrived at the office the second day with some notes that he made during his morning commute. Analysis does not just occur at the office. The comparison between observational data and electronic capture also revealed what we term "hidden gaps" or hidden activities. In a number of instances analysts were using or switching between both online and offline activities in close conjunction and coordination with each other. For example, the analyst might be reading from notes made on a hard copy document and using those as input to the report he is generating. Doing both activities more or less in parallel does contribute to a higher level of effort but does not appear to be easily captured electronically.

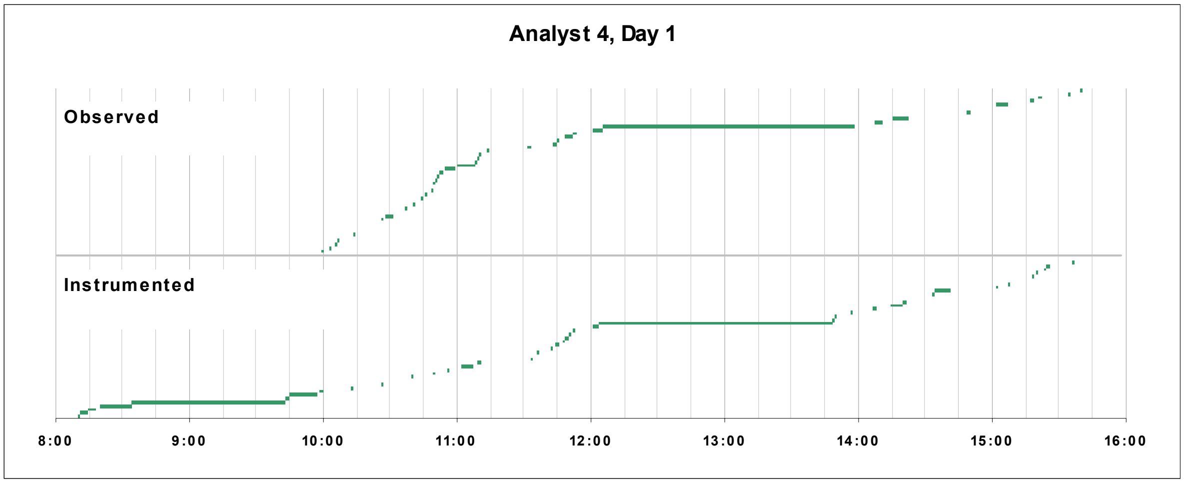

Figure 2: Periods of inactivity for Analyst 4 on Day 1

We also looked at the growth over time of the document that in both cases turned out to be the analysts' final report. Figure 3 shows the growth of the document for analyst 4 during the first day's activity. From these data it is quite clear that the analyst was engaged in report preparation even during the collection of initial information. For longer term tasks, the analysts report that they also collect information and enter it into a document as they did with this short term problem. We do not know yet if the report that results from a long term problem is the end result of being similarly modified and edited over time or whether the analyst makes a new document, using the original as a source of data for his report.

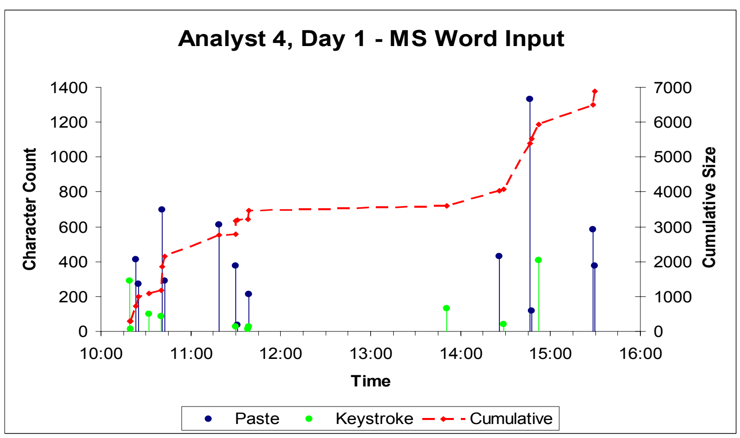

Figure 3: Growth of information/ time for Analyst 4, Day 1

Metrics and Measures

During the past year we have been examining a number of metrics to determine appropriate measures for assessing the effectiveness of a computerized support environment for intelligence analysts. As we want to measure this over long durations and/or in situations where no human observer will or can be present, we need to ensure that we can automatically collect appropriate and necessary data. From the results described here, one metric that is appealing is the amount of information/ effort/ time. That is, if an analyst working in a newly designed environment can obtain more useful information using less effort (fewer searches, fewer documents scanned) in the same amount of time or less time than in the current situation, the environment should be deemed effective. This is similar to Pirolli and Card's measure of information gain per unit time [7]. There is still, however, the issue of how to measure effort, in particular, cognitive effort.

Furthermore the problem of assessing cognitive effort appears to be compounded by the need to factor in the analysts' level of expertise or domain knowledge. Both analysts reported on here were domain experts and displayed some of the hallmarks of expert behavior [5]. They tended to perceive large meaningful patterns in their problem domain. In addition each tended to have good self-monitoring skills, quickly becoming aware of errors and making corrections. Since the level of cognitive effort will in part be a function of domain expertise it seems that expertise needs to be factored into the mix as well. Nonetheless, we think it should be possible to use estimates from the model human processor [2], or some other appropriate form of cognitive modeling, to compare cognitive efforts needed for data collection and organization, data exploration, and report generation in various tools. These estimates would need to be prepared and compared in a number of different situations involving different analysts, different tasks, and perhaps even different hardware configurations (such as multiple monitors, etc.) Human performance indicators for cognitive work could be used to corroborate the estimates.

However, we are still faced with the challenge of automating collection of the data required to compute information gained/ effort/ time. Our observational study has shown that analysis not only permeates data collection and report generation, it also does not stop on the desktop. Therefore, it seems that data collection mechanisms must go beyond the desktop as well. Devices such as tablet PCs, electronic ink pens, PDAs, and small voice recorders might be integrated into the current environment in order to allow more ubiquitous data collection. Experience sampling methods to capture data have been tried in ubiquitous computing environments [4]. Will analysts be willing to use devices such as tablet PCs to read and annotate documents as opposed to paper and pencil? It will probably depend upon whether the cognitive overhead associated with using the device is low relative to the perceived benefits to the analyst.

Conclusions

We used observational studies of two professional intelligence analysts to determine how much of the analytic process we could currently capture in an instrumented environment. We found that we could quite accurately capture lower level data such as time spent in applications, search terms used, web sites visited, and the number of documents viewed from search results. Offline activities such as telephone calls, meetings, reading offline, and annotating hard copy documents are not captured. More importantly, analysis itself is constantly taking place, during data collection and during report generation. Therefore it no longer seems appropriate to use the "inverted bathtub" measure. That is, it does not seem that leaving more time for analysis per se as opposed to data collection and report generation is the desired goal. Our measure of amount of information gained/ effort expended allows for analysis to occur throughout the data collection phase but by lowering the effort extended for data collection, more of the cognitive process can be devoted to assimilating the new information into the evolving analysis. To accurately compute this, however, we need to find ways of automatically collecting data beyond the desktop. This will involve incorporating new devices into the current analysis environment and may mean being more intrusive in data collection efforts. This will need to be done with great care so as not to overburden the analyst.

References

1. ARDA NIMD home page, http://www.ic-arda.org/Novel_Intelligence/, accessed Dec. 28, 2003
2. Card, S.K., Moran, T.P., & Newell, A. (1983). The psychology of human-computer interaction. Hillsdale, N.J., Lawrence Erlbaum Associates.
3. CIA Factbook, http://www.odci.gov/cia/publications/facttell/intelligence_cycle.html, accessed December 28, 2003.
4. Consolvo, S. & Walker, M. (2003). Using the Experience Sampling Method to Evaluate Ubicomp Applications, Pervasive Computing, Vol 2(2). 24-31.
5. Glaser, R. & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser & M. J. Farr (Eds.). The nature of expertise. Hillsdale, NJ: Erlbaum.
6. Heuer, R. (1999). Psychology of Intelligence Analysis. CIA Center for the Study of Intelligence.
7. Pirolli, P. & Card, S. (1995). Information foraging in information access environments. Proceedings of CHI'95. ACM Press. 51-58.
8. Wilkins, D. (2002). The Bathtub Curve and Product Failure Behavior. Reliability HotWire, Issue 21, November. www.weibull.com/hotwire/issue21/hottopics21.htm. Accessed December 29, 2003.