Metrics and methodologies for evaluating technologies for intelligence analysts[1]

Emile Morse, Michelle Potts Steves, Jean Scholtz
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, Maryland 20899, USA
{emile.morse, michelle.steves, jean.scholtz}@nist.gov

Keywords: software evaluation methodologies, user-centered evaluation, strategic intelligence

Abstract

In this paper we discuss the evaluation methodologies and metrics we have developed for ARDA's Novel Intelligence for Massive Data (NIMD) program. We discuss the requirements for developing methods and metrics in a situation where software components that were to be tested were in very early stages of development and where investigators who might be on the leading edge with respect to their technology were novices with respect to evaluation. Additionally, we discuss how our process of evaluation design is evolving as we gain experience with metrics and measures that are obtainable, yet have some value as indicators of future software performance in the field.

1. Introduction

Evaluation is a key component of the NIMD program (http://www.ic-arda.org/Novel_Intelligence/). We at NIST have worked on developing methods that allow software development projects to be tested early and often. NIST's role in evaluation for NIMD is two-fold: 1. Work with NIMD researchers to develop methods to test their software products and, 2. Develop common metrics for each of the NIMD research areas. The hope is that metrics developed during one evaluation can be re-used by subsequent evaluations within the same area and used in the longer term to compare the analytic products and processes with and without NIMD tools.

2. Methodologies

In the past year, we employed two user-centered evaluation methods – heuristic review and user testing. Two NIMD projects were subjected to a heuristic evaluation. User testing was applied to 9 pieces of software from 7 different projects. Table 1 lists the NIMD research areas and shows the number of projects that have been evaluated in each research area compared with the total number of projects in that area.

Table 1: Number of projects evaluated by research area

NIMD Research Areas	evaluated/total
Modeling Analysts and Analytical Processes	2/5
Prior and Tacit Knowledge	1/4
Hypothesis Generation and Tracking	3/7
Massive Data	1/5
Human Information Interaction (HII)	2/4

The major components of a user-centered evaluation are the subjects, their tasks, and the data collection with which the subjects work. In addition, methods for collecting data from the subject and from the subject's interaction with the system need to be considered. We used the Glass Box software (Cowley, Nowell, and Scholtz, 2005) to facilitate data collection. The Glass Box collects keystroke data, information about queries that are made, web pages that are accessed, versions of documents that analysts create, annotation data, video screen capture and audio data from the subjects.

For each evaluation, three to six subjects were recruited from the Navy Reserves. They performed intelligence analysis as their reserve duty and had from 1 to 19 years of experience. Each system is unique and requires a specialized experimental design. The following are the designs we used in these 9 evaluations:

- Analysts generated hypotheses given a number of pieces of information. They compared their own hypotheses to the ones generated by the software.
- Analysts rated the relevance of query results submitted to the software system with user modeling and a baseline system.
- Analysts solved text book problems using research software.
- Analysts solved a scenario by hand and compared it to a system-generated solution.
- Analysts evaluated a system's intermediate representation for completeness and correctness.

[1] This work was supported the ARDA NIMD Program.

- Analysts solved problems with and without the use of the system.
- Analysts worked with two versions of the system – with and without a key feature.

Data collections ranged from the Web to custom data sets to collections such as those distributed by the Center for Nonproliferation Studies (CNS).

It is interesting to note the variety of comparisons that these studies employed. Although the evaluations were primarily exploratory in nature and we knew that we would not be striving for statistical types of outcomes, the studies incorporated comparisons many of which could be used for evaluations in the future.

3. Metrics

The second focus of our work has been to develop metrics for evaluating NIMD systems. To date, this has mostly been a bottom-up effort, i.e., specifying metrics and measures for each evaluation that are specific to that evaluation. Table 2 shows the metrics and measures that were developed for one of the NIMD research areas – Human Information Interaction. In reviewing the metrics and measures used in our evaluations, some commonalities have surfaced which led to the categories shown in Table 2. Similar sets of metrics have been developed for each of the NIMD research areas.

Table 2: Metrics for Human Information Interaction research area

Efficiency
- time/search
- time/document read

Effort
- # documents accessed
- # documents read
- document growth rate
- document growth type (cut/paste vs. typing)

Accuracy
- Evidence used in analysis
- Number of hypotheses considered
- Average system rank of documents viewed

Confidence
- User confidence ratings of findings

Answer/Report Quality
- Quality of report
- Ranking of report

Cognitive workload
- Cognitive workload ratings (NASA TLX; Hart & Staveland, 1988)

In addition to the above data, user questionnaire data were collected with respect to user demographics and to place the user's experience with the tool being evaluated in a context relative to the analyst's work environment. Finally, it is pertinent to note that many of these metrics were gathered by analyzing system logs and observation notes. At times observers captured timing information. At other times, Glass Box log data was mined for the timing information. Developers logged information from their systems as well that allowed us to capture additional data.

4. Next Steps

As we move forward in gaining experience in evaluation of systems supporting the intelligence community and as those systems mature, new directions are emerging. As systems mature, we need to start to consider variables such as the amount of time needed for analysts to become familiar with the system; for systems that feature user modeling, the amount of time needed until the system becomes useful for the analyst; and to facilitate comparisons across systems and over time we need to consider the issue of equivalence of complexity of tasks.

We plan to use a *metrics model*, i.e., framework, developed at NIST for organizing the many and varied metrics and measures and their associated context [Scholtz & Steves 2004, Steves & Scholtz 2005]. The metrics model provides a top-down approach for specifying system goal-directed software evaluations. Use of the metrics model in future evaluations has several envisioned benefits, namely: more re-use of metrics and measures for similar evaluations and the possibility of comparison of like-structured evaluations.

We have completed user-centered evaluations for 7 NIMD software projects for a total of 9 software tools. A wide variety of testing methodologies were employed to meet the needs of the researchers – knowing how their software performs in the hands of real analysts – and the needs of NIST – finding out what kinds of metrics have promise in evaluating NIMD systems. We have identified metrics that work across research areas and some that work well within an area. Although the methods and metrics were developed specifically for the NIMD software research tools, we think that the methodologies and metrics are also applicable to other research. We are currently applying a number of these to other software tools designed to help intelligence analysts.

References

Cowley, P., Nowell, L., and Scholtz, J. 2005. Glass Box: An Instrumented Infrastructure for Supporting Human Interaction with Information. HICSS 38. Jan 3-6. Hawaii

Hart, S.G. and Staveland, L. E. 1988. Development of a NASA-TLX (Task load index): Results of empirical and theoretical research. Hancock, P. and Meshkati, N. (eds.), *Human Mental Workload*, Amsterdam: North-Holland. pp. 139-183.

Scholtz, J. and Steves, M. 2004. A Framework for Real-World Software System Evaluations. Proceedings of Computer-supported Cooperative Work, 600-603.