A Study on Search Engine Use By Intelligence Analysts

Don Libes <don.libes@nist.gov>
Emile Morse <emile.morse@nist.gov>
National Institute of Standards and Technology
Gaithersburg, Maryland, United States

Jean Scholtz <jean.scholtz@pnl.com>
Pacific Northwest National Laboratory
Richland, Washington, United States

ABSTRACT

This work reports on search engine use by intelligence analysts, generally considered experts at the task of searching for information. Intelligence analysts share a degree of common training and understanding while having widely differing backgrounds. This includes their approach to software tool use – analysts show signs of a common understanding while also displaying uniquely personal differences in their investigative approaches. We present some of these observations from studying their use of an open-source search engine. We also report on the ability to track paths of analyst investigation by studying search engine activity.

KEYWORDS

Data mining; Intelligence analysts; Search engines

INTRODUCTION

This work reports on search engine use by intelligence analysts, generally considered expert at the task of searching. By studying their use of search engines in an open-source environment, we have been able to make a number of useful observations and create new software to assist in further research of such analysts.

In order to better understand these observations, it is important to understand the context for this work. The central focus of our work on the Novel Intelligence for Massive Data (NIMD) program is to create metrics for the evaluation of software and practices for intelligence analysts. During the creation of such metrics, we have necessarily studied analysts' current practices.[1][2]

While analysts have widely differing backgrounds and working styles in certain respects, we have observed commonalities that are quite interesting.

We believe these observations are likely to be useful in several respects. First, researchers who are building next-generation software for intelligence analysts need a better understanding of their users. Second, trainers of intelligence analysts can use this material to better adapt their teaching materials and methods. Third, management can get a better understanding of the practices of their analysts. Our own work in the development of software metrics is aided by a better understanding of analyst behavior. Finally, analysts themselves may use this material to gain increased efficiencies in their own practices.

This report is based on data collected from analysts over several years. Specific figures are provided for four analysts during a five-week period. These analysts were conducting open-source analysis and their interactions with various software tools were captured to provide NIMD researchers with data about analytic processes.

Our work included the building of a software system called "Degoo" to provide automated analyses of traces of search engines. [3] The name is historical. Degoo first analyzed Google queries.[1] It has since been expanded to handle other search engines (Ask, Yahoo, Wikipedia, etc.). [4][5][6]

1. Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.

Raw data was collected using the Glass Box, instrumentation that logs all interaction a user makes with a computer. This includes low-level activity such as keystrokes and mouse motion. A variety of higher-level interpretation is done to detect cuts/pastes, menu interactions such as file opens/saves, and prints.[7]

This can be used, for example, to deduce the quality of query responses. We can infer the utility of a search by counting the number of documents printed or saved. Analysts are also able to explicitly leave comments ("annotations") explaining what they were doing by using features of the Glass Box. They can use these annotations to describe analytic strategies, tell us about problems with software, or document offline activities. However, we want to reduce our intrusive interactions with analysts as much as possible. As the name "Glass Box" suggests, logging and analysis is to be as transparent as possible.

DEMOGRAPHICS

Four analysts were studied: one senior analyst, one mid-level analyst, and two junior analysts. For historical reasons, analysts were numbered beginning with 4.

- Analyst 4 worked on six tasks related to Mid-East governmental issues.
- Analyst 5 worked on seven tasks related to cybersecurity, cyberterrorism, and missile defense.
- Analyst 6 worked on four tasks related to biologic and genetic research.
- Analyst 7 worked on nine tasks related to Mid and Far-East medical and weapons of mass destruction issues.

DATA ANALYSIS

802 search engine queries were made and examined. All but 7 queries were from google.com. The remaining queries were from:

- wikipedia.org
- search.yahoo.com
- ask.com
- nti.org

Due to the overwhelming reliance on Google by the analysts, we concentrated our investigations on Google queries. Parsimoniously, Google makes an excellent candidate for study. It is relatively easy to track parts of its operation. For example, the Glass Box interposes itself as a proxy so that a complete log of all http requests is recorded.[8]

Furthermore, Google is used by all the open-source analysts we have observed. Some analysts use Google more than others and in different ways, making it an interesting source of comparisons. But in some sense, Google is perhaps the most common way to access unstructured data in the open literature.

- Analyst 4 made 71 Google queries.
- Analyst 5 made 127 Google queries.
- Analyst 6 made 162 Google queries.
- Analyst 7 made 436 Google queries.

SEARCH ENGINE CAPABILITIES

Google provides simple keyword matching and its primary interface suggests little more than that. However, Google also offers many other tools. For example, Google will search for pages that:

- contain all the search terms
- contain the exact phrase
- contain at least one of the terms
- do not contain any of the terms
- are written in a certain language
- are in a certain file format
- have been updated within a certain period
- contain numbers within a certain range
- within a certain domain or website
- contain synonyms of terms

Some of these are provided using search operators, such as putting a "+" prefix on a term. Others capabilities are available through an "Advanced Search Page".

We found analysts used only a fraction of the capabilities that Google offered. Specifically, we found and studied the analysts use of:

- basic term search
- term negation
- search by reference to similar pages
- spelling-corrected search
- "I feel lucky" search

COMMON OBSERVATIONS

Here is an example stream of Google queries. Each line represents a single query. The letters (A, B, C, and so on) are placeholders for real keywords.

A
B
C
C D

C D E
C D F
X Y Z

In this stream, the analyst starts by searching for A. Getting inappropriate results, in the next query, A disappears and B is tried. Then C. Finally, the analyst gets responses but too many. So the analyst adds on new keywords one at a time to end up with C D E.

This is a fairly typical sequence. From here, an analyst might go on to do different things, shown in the last two lines. In the first of the two lines, one keyword (E) has been deleted and exchanged with another (F).

In the final query, all the keywords are different. This might happen because the previous query has led down a dead-end. Alternatively, the change signals an entirely new area of investigation.

In reality, it is unusual to find such a precisely intuitive line of querying as we have described here – various complications nearly always arise. As an example, figure 1 illustrates an actual sequence produced by Degoo for analyst 6. Notice that the first two queries are identical. Why?

Repetitions come about for a variety of reasons. For example, an analyst may have intended only to search for the first keyword but the browser software may have auto-filled the remainder as a "helpful" shortcut. Alternatively, the analyst may have started a new browser session (perhaps interrupted by lunch) and executed the same search intentionally – not having saved the results from the last time. It is even possible that the analyst simply may not have been paying attention or wanted to see the results in a new window or any of a number of other explanations. (Window tracking and time analysis are used to distinguish these situations but further discussion is beyond the scope of this paper.)

This is a good example of an unexpected outcome. Because all Google searches take much less time than it would require to save and find the results of a previous query, analysts do more searching than they technically need to do. Indeed, we see much evidence of seemingly redundant searching. In addition to the other reasons we have mentioned, it is also important to recognize that the responses to some queries change over time and analysts sometimes want this. Such queries can generally be recognized by a more evenly distributed pattern of identical queries.

Alternatives

If Google deduces that a keyword is likely a misspelling, it will offer an alternative. This is an aid that is useful and gives analysts the impression that Google has some (albeit limited) understanding.

However, analysts do not recognize that Google has no special handling for some query attributes. For example, Google ignores letter case. Therefore, *US* will match both *US* and *us*. This is also true even if the string is in quotes. So *"US magazine"* will also match "*Us magazine*".

All analysts showed evidence of this. For example, 33 of analyst 4's 77 queries used mixed case. The other analysts had successively more queries with comparable percentages of mixed case.

These observations are the tip of the iceberg. Google has a large number of rules that are intuitive to some and surprising to others. Traces of Google queries illustrate that.

Common Words

Some analysts enter queries using words such as "an", "the", "of", and so on. Other analysts do not, evidently thinking these words are inconsequential and that omitting them saves keyboard entry time.

However, Google is sensitive to such words and returns different results depending on their presence and absence.

dna	databa									
dna	databa									
dna	databa	bush	propos							
dna	databa	bush	propos	crimin						
dna	databa	bush			fbi					
dna	databa	bush			fbi					
						status	of	bills		
						status	of	bills	united	states

Figure 1: A simple query trace but with repetitions.

This confusion over what Google is or is not sensitive to is not surprising because Google is insensitive to other information that analysts (and most users in general) think significant. As mentioned earlier, Google pays no attention to letter case even when quoted. Yet analyst logs shows that all the analysts entered case-specific queries, evidently not realizing that case is irrelevant.

In fact, case is relevant to Google but not in an intuitive way. Rather, Google uses case to signal such things such as boolean constructions (e.g., OR) and other meta-words. However, none of our analysts used uppercase metawords.

These types of "casual discrepancies" are typical of many user interfaces and cause unfortunate problems when users do not realize their existence. User-interface designers would do well to take this in to account when creating such user interfaces.

UNIQUE OBSERVATIONS

There is a wide disparity in how different analysts use Google.

Simple But Unused Power Tools

As an example of the disparity in usage, only one analyst was found to use Google's negation mechanism – the ability to make a query return results that did not match a particular term. This was true whether the negation was in-line or via the (Advanced Search) form.

We had expected negation to be a commonly used refinement mechanism since it is not only intuitive but extremely valuable when sifting through queries that are otherwise too productive. Indeed, the premise underlying the NIMD work is that the glut of information is one of the problems preventing analysts from being more productive. Yet if our studies are any indication, some of the simplest tools to control this massive amount are not being used.

This should serve as a lesson to software designers – Make sure that you understand how users will use your software. It is often different than how you might imagine during the design cycle. Focus on why users fail to take advantage of simpler ideas before moving on to the sophisticated ones.

Another example that illustrates the lack of analysts' reliance on power tools is the relationship mechanism offered by Google. After a successful search, this is triggered by clicking on the "Similar pages" link. The idea is that once Google has returned useful information, Google has a better idea of what the user is looking for.

In the example shown in figure 2, the Google query on "terrorism" returns 52,400,000 documents, a rather staggering number. Assuming the analyst is indeed happy with that link and wants to see more pages like that one, choosing "Similar pages" narrows the number of documents down to just 25, a reduction of over 2000%!

In a similar vein, only one analyst was found to use the "I Feel Lucky" mechanism. This mechanism skips the intermediate step of offering the analyst ten choices and instead jumps directly to the first of those ten. We observed this used not as a matter of luck as the name implies but as a shortcut – when the analyst knew in advance what search terms were necessary to bring up a perfect hit on a page.

As an example, an unskilled analyst interested in finding out what the term "smartcard" meant might enter it into Google only to be deluged with dozens of sites for smartcard vendors. However, a skilled analyst might more quickly enter "smartcard wikipedia" and click the Lucky button to skip right to an encyclopedic explanation of the subject.

Meta Words

Earlier, we mentioned the existence of meta words. We found even the simplest ones (such as OR to indicate a match of either of two search terms) were never used.

Figure 2: Google's relationship mechanism pointed to by arrow.

In fact, Google has a rich vocabulary of such meta words that direct the search. However, if it required hand-entry, such mechanisms were never used. In contrast, Google offers a search form which hides some of the details. Using this form, the meta words were used, albeit only a few and rarely.

TRACKING AREAS OF INVESTIGATIONS

One of the motivating aspects of our work was tracking the different lines ("paths") of investigation being explored by analysts. We are interested in such questions as: When and why did analysts shift their focus from one subject to another? When assigned multiple tasks, did analysts merely do the task with the nearest deadline? Can we tell when analysts are working on tasks that were never assigned explicitly?

To help this research, the Glass Box supports the ability to track when analysts shift their focus from one task to another task. However, the tracking requires an explicit interaction by the analyst – to pull down a menu and identify the task. Analysts do not always do this, whether intentionally or not. Thus we are interested in knowing whether this information can be inferred or even how closely an analyst's notion of a particular task corresponds to reality.

After all, an analyst may consider two tasks too closely related to separate for administrative tracking purposes. In addition, analysts generally are highly specialized in their particular fields of inquiry already. Or one task may be a longer-term version of a shorter-term task. In yet a third scenario, analysts may not even be consciously aware that they are pursuing a new line of reasoning that may not even have been assigned explicitly because it was unknown heretofore.

For such reasons, it is interesting to see how automated reasoning tools may view the analyst's focus.

Semi-Automatic Algorithms

Our tracking algorithms are presently a mix of automatic and manual techniques.

Most of the work is done automatically by relating queries that have words in common. For example, the query "terrorist communication methods" is presumably related to the query "terrorist communication encryption" even though "methods" was dropped from the query and "encryption" was added.

Using this idea, a transitive closure is used to produce a web of related queries. An ontology is used to establish relationships between keywords that are not otherwise obviously related. For example, the words "stealth" in one query and "covert" in another query might be reason enough to relate them. More commonly, alphabetically close words such as "cryptography" and "cryptoprocessor" need to be related.

Figure 3 is a snapshot of a section of Degoo's ontology window while Degoo is being used to analyze the activity of analyst 5.

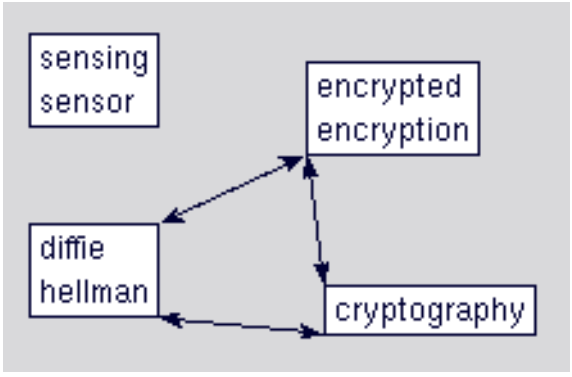

Figure 3: A snapshot of a small example of Degoo's guiding ontology relevant to guide the analysis of a set of queries of analyst 5. Synonymous terms appear in the same box. Related terms are connected by arrows.

At present, the ontology is custom built and manually maintained. It can be changed on the fly while running the Degoo software to accommodate the realities of the relationships. We envision building in support for extant ontologies in the future to help with this.

The ontological support also includes some heuristics for word matching and word exclusion. For example, certain keywords such as "is", "the", and "2005" should generally be ignored while trying to derive relationships between queries. While terms such as "2005" are clearly useful, it is surprising how frequently analysts use words that are so common as to be unhelpful.

Figure 4 shows the result of the search path tracking algorithm subsequent to semantic analysis and transitive closure on tasks performed by analyst 4. Each line shows the keywords used in a single query. The lines appear in chronological order. The analysis shows evidence of three paths, illustrated by different background colors. (A larger window would have

shown that Degoo found all six path performed by analyst 4. However, our screen snapshot software restricted us to the size of the physical display being used.)

SUGGESTIONS FOR TOOL BUILDERS

Although this was not the focus of our work, we can make several observations that should be observed by tool builders in the intelligence community. (Some of these suggestions may apply more widely.)

- If you target software so that it *can* be used without training, you virtually guarantee people *will* never get training, no matter how much that training might help.
- Make feedback painless. Google's "did you mean" is easily ignorable. Contrast this with the Microsoft paperclip.[9]

SUGGESTIONS FOR ANALYST MANAGEMENT

- Software that appears to be intuitively obvious to use may be, but only at a superficial level. Commit funds and time for training to increase the usefulness of such tools.
- Testing software products can be worthwhile. However significant resources are necessary to do testing right.

SUGGESTIONS FOR ANALYSTS

- Be careful of embracing intuitive interfaces with the idea that your job will be easier with less training

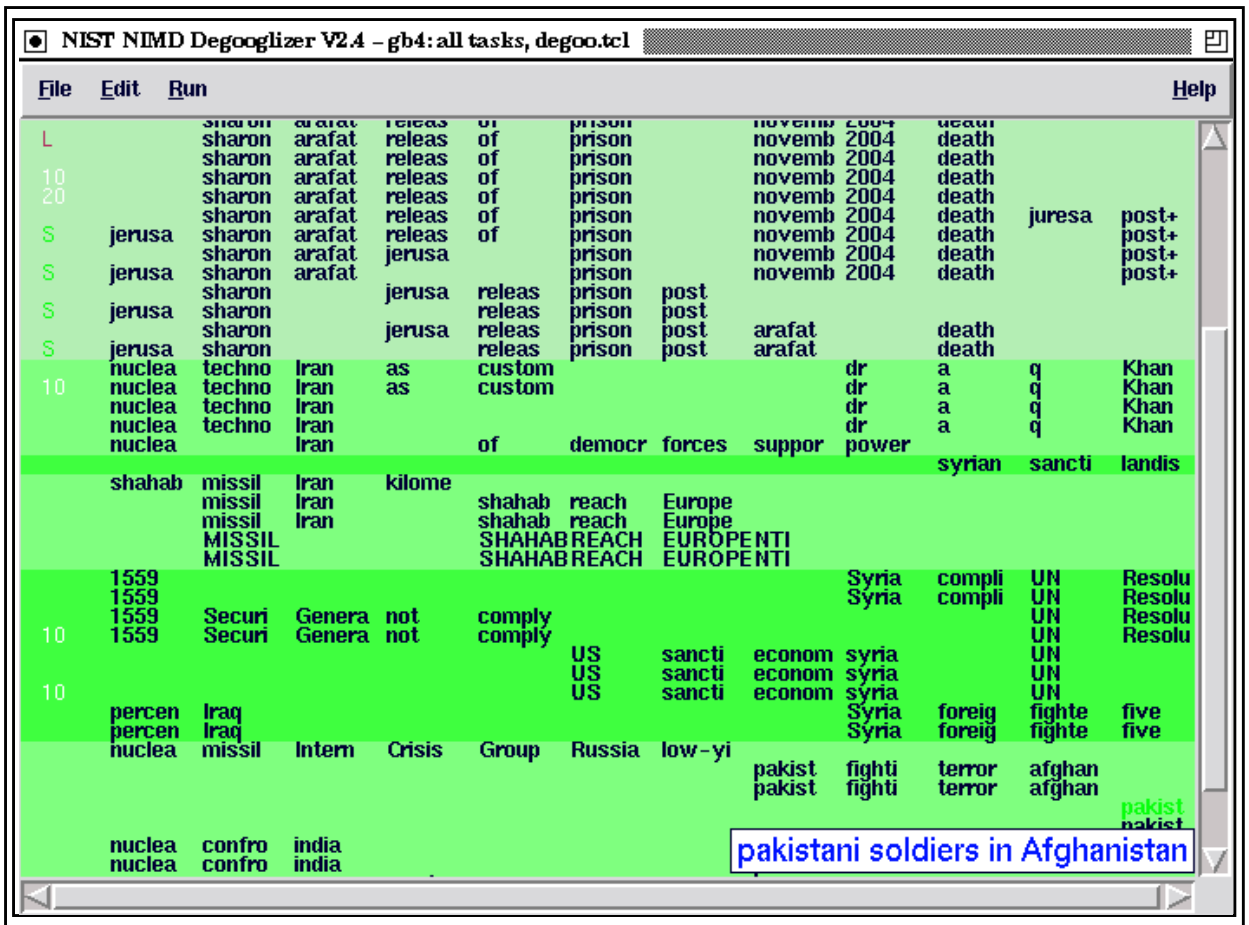

Figure 4: Output of search path analysis showing three different paths of inquiry (indicated by three different color backgrounds) performed by analyst 4. Each line represents a single query; each column a single keyword. Keywords are abbreviated to save space. The popup at bottom right shows one of the complete keywords as the mouse rolls across the columns. Notice how the second path starts with queries about Iran and ultimately leads the analyst to queries about Russia and "pakastani soldiers in Afghanistan." The left-hand column includes additional insights. For example, the L in the upper-left hand corner indicates the query was generated from the "I Feel Lucky." button.

and more efficiency. The two are not necessarily connected.

- Intuitive interfaces are often just a tip of the iceberg, hiding more sophisticated and powerful mechanisms that are worth spending time to learn.

FUTURE WORK

We suffer from the very same problem that NIMD was borne of – a glut of information requiring that we necessarily ignore some of it to focus on anything. We recognize that we have just touched the tip of the iceberg. We would like to explore these areas more deeply as well as many other related areas.

We particularly think it useful for future studies to "close the loop" and examine how analysts make use of the query responses and how that guides their future queries. A second line of research is to apply the results described in this paper to other tools that analysts use.

Finally, we also would like to incorporate additional ontological and semantic reasoning. Query patterns encountered suggest that analysts would benefit from increased aids including semantic analysis, user modelling, and more intensively graphical interfaces.

CONCLUSIONS

We found that it was possible to do path tracking of intelligence analysts who were using Google. This is particularly useful to managing analyst task assignments and is useful for additional analysis of analyst activity.

We also found that Google's intuitive interface has interesting side-effects ranging from both increased efficiency to decreased inefficiency but for different reasons. We have described several ideas to enhance productivity. We believe these are applicable not just to intelligence analysts but to search-engine users in general.

Of equal significance is the creation of tools more specialized to the analyst. While open-source search engines are popular, our research suggests that analysts would benefit frequently from tools that do semantic analysis, that develop models of the analyst goals, present output graphically and provide graphical control in the user interface. Such tools are under development already and our observations of how analysts use search engines with seemingly intuitive user interfaces have made us aware of the training that has to be done for these new tools to ensure that analysts take full advantage of their capabilities.

Although we have focused primarily on Google in this paper, the observations described herein are applicable to most other search engines as well as a wide-variety of other software tools and user interfaces.

ACKNOWLEDGMENTS

Thanks to Michelle Steves, Al Jones, and Sharon Kemmerer for comments and discussion that improved the paper.

REFERENCES

[1] Novel Intelligence from Massive Data, http://www.sourcewatch.org/index.php?title=Novel_Intelligence_from_Massive_Data, July 2006.

[2] J. Bodnar, *Warning Analysis for the Information Age: Rethinking the Intelligence Process* (Washington DC: Center for Strategic Intelligence Research, December 2003).

[3] Placeholder for author during refereeing, Degoo, Available on request to placeholder@placeholder, July 2006.

[4] Ask, http://www.ask.com, 2006.

[5] Wikipedia, http://www.wikipedia.org, 2006.

[6] Yahoo, http://www.yahoo.com, 2006.

[7] P. Cowley, L. Nowell, and J. Scholtz, Glass Box: An Instrumented Infrastructure for Supporting Human Interaction with Information, PNWD-SA-6555, 2005.

[8] Google, http://www.google.com, 2006.

[9] E. Leuning, Microsoft tool "Clippy" gets pink slip, http://news.com.com/2100-1001-255671.html, April 11, 2001.

[10] Disruptive Technology Office (DTO), http://en.wikipedia.org/wiki/Disruptive_Technology_Office, July 2006.