SAND2016-1652C

Improving Analysis and Decision-Making through Intelligent Web Crawling

Jonathan T. McClain[1], Glory Emmanuel Aviña, Derek Trumbo, & Robert Kittinger

[1]Corresponding Author, Sandia National Laboratories, Albuquerque, New Mexico [jtmccl@sandia.gov]

Abstract. Analysts across national security domains are required to sift through large amounts of data to find and compile relevant information in a form that enables decision makers to take action in high-consequence scenarios. However, even the most experienced analysts are unable to be 100% consistent and accurate based on the entire dataset, unbiased towards familiar documentation, and are unable to synthesize and process large amounts of information in a small amount of time. Sandia National Laboratories has attempted to solve this problem by developing an intelligent web crawler called Huntsman. Huntsman acts as a personal research assistant by browsing the internet or offline datasets in a way similar to the human search process, only much faster (millions of documents per day), by submitting queries to search engines and assessing the usefulness of page results through analysis of full-page content with a suite of text analytics. This paper will discuss Huntsman's capability to both mirror and enhance human analysts using intelligent web crawling with analysts-in-the-loop. The goal is to demonstrate how weaknesses in human cognitive processing can be compensated for by fusing human processes with text analytics and web crawling systems, which ultimately reduces analysts' cognitive burden and increases mission effectiveness.

Keywords: text analytics, intelligent web crawling, decision making, cognitive consistency

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000, Sandia Report 2015-1424C. Approved for public release; further dissemination unlimited. This research was funded in part or whole by an Interagency Agreement between the Transportation Security Administration and the Department of Energy.

1 The Challenge of Data Analysis

While the prevalence of easily accessible information via the internet and large databases has allowed for unprecedented advances in societal knowledge, the sheer volume of data available leads to difficulties in locating the correct information that is relevant to a task at hand (i.e. finding the needle in the haystack). This sifting process is most commonly accomplished today using search engines (e.g., Google®) by submitting a single query and iteratively visiting results in a single list to determine whether the supplied result contains relevant information or represents a false positive. This process continues iteratively through multiple queries until the information required has been found or the human analyst gives up. While such a process is useful, the overall approach itself suffers from a number of problems; 1) the analyst is required to possess a moderate understanding of the subject matter being sought; 2) the analyst is limited by the query interface provided and must possess astute abilities in constructing queries with a few words to seek out that subject matter; 3) the analyst is limited by the fact that a single prioritized list is presented based on an unknown underlying search algorithm; 4) in many cases, search algorithm results are tailored either to the global mean, or tailored to the analyst, both of which may be undesirable when searching for obscure and little known information; 5) the analyst is limited to that information which the search engine has deemed worthy of indexing, also based on global demand (e.g., Google® only indexes a small fraction of the known internet)[1].

1.1 The Challenge for the Analyst

Even the most experienced analysts are unable to be completely consistent and accurate when sifting through large amounts of information. A single analyst faces a number of cognitive hindrances. An analyst will use heuristics, such as scanning for words they have determined to be relevant, in order to gauge information importance[2]. However, this method is inconsistent. At the start of the analysis process, an analyst can decide a document is relevant because of the words in a piece of text, but later on, after they have been sifting through information, they decide a similar piece of text is not relevant because their notion of what is relevant has matured. Similarly, relevancy is based on what the analyst knows to be important and therefore is biased to their limited knowledge base on the subject of interest. A single analyst must also spend large amounts of time examining and filtering large amounts of documentation, and even then he or she is unable to synthesize and process all of the data, especially when there is a limited amount of time to make important decisions[3].

The cognitive hindrances increase for a team of multiple analysts. Between analysts, there are different heuristics and various strategies for finding information. The amount of time spent searching is multiplied by how many analysts are on a team,

[1] http://www.webanalyticsworld.net/2010/11/google-indexes-only-0004-of-all-data-on.html
[2] Goldstein & Gigerenzer, 2002.
[3] Pope, Ziebland, & Mays, 2000.

which can make searches for relevant information expensive. In addition, biases towards determining what information is relevant increase because of differences in experience, knowledge base, and perspectives[4]. Conflict may also arise if there are conflicting opinions of documentation relevance. Finally, if an automated method for tracking information examined is not used, then analysts may have overlap in the material they have covered[5].

1.2 The Challenge for the Decision Maker

Ultimately, when a single analyst or team of analysts present the information they have determined to be relevant and their assessment of it, it is inevitable that they have not located all relevant information. Therefore, conclusions based on analysts' information are automatically biased, limited in scope, and skewed to the cognitive perspective of the analysts. This creates a challenge for decision-makers because they need to be able to justify their conclusions. In order to make defensible decisions, the decision-maker needs to have access to analyses and conclusions that are accurate, quantitative, justifiable, and thorough. This holistic assessment provides the pathway for decision-makers to not only make decisions, but also anticipate and respond to potential issues that the data alludes to as well as predict how and why situations may evolve.

This need for complete data does not point decision-makers to a fully automated system. Such a system could not spot the nuances in the data that a human analyst so naturally does. Instead, decision-makers need to keep the human-in-the-loop to leverage analysts' intuition, ability to calculate possible options in connection to the scenario at hand, and create a continuous pathway from the data to solutions.

Overall, the decision-maker as well as the analyst needs to reduce the amount of data processed by humans and therefore cognitive load to increase effectiveness, accuracy, and speed.

1.3 The Problem with Search Engines

When searching for relevant information on the open web, a primary question when presented with intelligent web crawling is, "What about search engines?" This question stems from an underlying assumption that search engines are enough to satisfy analysts' needs. However, if one thinks about this assumption in a deeper way, it becomes evident that search engines, even the best of them, are not the end-all solution for sifting through large datasets for relevant information.

Imagine you as an analyst are going to use a search engine to find information on a topic of interest such as the spread and impact of your academic thesis. If you search for the title of your thesis to find relevant information related to your thesis topic, you would receive a single list of webpages that have the words from your title on them. Your search will probably return your institution's academic repository and

[4] Marchionini, 1997.
[5] Howard et al., 2009.

possibly the journal where you may have published your thesis. From there, the list may be your personal website, and then from there a list of other websites. You do not really know why the search engine listed the other webpages except that there are a few keywords matching your thesis title on the webpage. Your job is now to sift through the results, probably going through the list of pages top-down to determine what is actually relevant to the question you are asking. If you have multi-dimensional parameters (e.g., wanting to find related publications and individuals who have quoted your work), this list of search results will not efficiently respond to both parameters. You will probably have to do multiple queries to answer each of these parameters. Analysts quickly find that a single metric such as a search engines' list of results is not enough to ascertain the quality of the results you are looking for.

Another problem with search engines is the lack of transparency. The reasoning for why a search engine presented a list of search results can be partially or completely hidden from the user.

Search engines also transform results according to user's location, personality, past purchasing and browsing behaviors, global and/or local trends, and are influenced by search engine optimization by third parties. Results are also dependent on the parameterization of the search engines' crawlers and search engines make tradeoffs to crawl/index less to save money. Furthermore, you do not have access to the full content available on the internet. The actual size of the internet already has made effective indexing infeasible[6] and Google specifically only indexes a small fraction (~.004% as of 2010) of the internet[1].

2 Huntsman

Sandia National Laboratories has attempted to solve the challenges faced by analysts and decision-makers by developing an intelligent web crawler called Huntsman. The use of web crawling and text analytics helps to both imitate as well as enhance human analysts by using text algorithms to develop consistent metrics to search and analyze large datasets. The search also eliminates bias, is parallel across computing machines, and returns the best matched information relative to all the data searched. Intelligent web crawling finds the most pertinent information and quantitatively pushes it to the forefront of the analysts' attention. This way, analysts are still in the loop to examine a smaller, more relevant dataset and validate findings.

Huntsman acts as a personal research assistant by browsing the internet or offline datasets in a way similar to the human search process, only much faster (millions of pages per day), by submitting queries to search engines and assessing the usefulness of page results by analyzing full-page content with a suite of text analytics. Huntsman uses the results of these analyses to order future downloads, allowing it to hone in on important information quickly. In this way, Huntsman provides a triage of information through analyzing the full content of each document to assess relevance to the task at hand. Upon completion, Huntsman provides various subsections of the

[6] Chakrabarti, S., Berg, M. v. d., & Dom, B. (1999) and M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork (2000).

data, based on the various analytics performed, to a human analyst, allowing them to focus only on the most useful information at hand.

2.1 Background

Intelligent or focused crawling is not a particularly well researched topic. Chakrabarti and Dom. (1999) first described a focused crawler that utilized a classifier to identify relevant documents, and a distiller to identify nodes which access several relevant documents within a few links. Zeinalipour-Yazti and Dikaiakos (2002) describes the idea of using web crawlers as middleware for users to gather relevant content based on a user profile.

Where Huntsman differs from these previous approaches is in its focus the human in the loop. Huntsman focuses on leveraging the humans' abilities in pattern matching and intuition, while eliminating tasks in which the human does not excel by removing the burden of mentally processing large amounts of data, the bulk of which is not relevant to the task at hand. Another area in which Huntsman differs from other approaches is in comprehensiveness. When data is processed with Huntsman, the analyst and the decision maker have much more confidence that all relevant information has been taken into account as part of the analysis.

2.2 How Huntsman Works

Unlike regular keyword-based analysis using search engines, intelligent web crawling helps alleviate analysts' tasks that are most subject to cognitive hindrances (biases, inconsistency, etc.) and keep analysts in the loop where they are most critical (intuitive decision-making, option calculating, etc.).

The process of using Huntsman begins with crawl parameterization. This includes identifying known documents and keywords and phrases of interest. The documents of interest are then passed through a suite of text modeling tools to create signatures that target both generally relevant, as well as specific content. Keywords and phrases are used to enhance these signatures by scaling their influence based on overall document relevance.

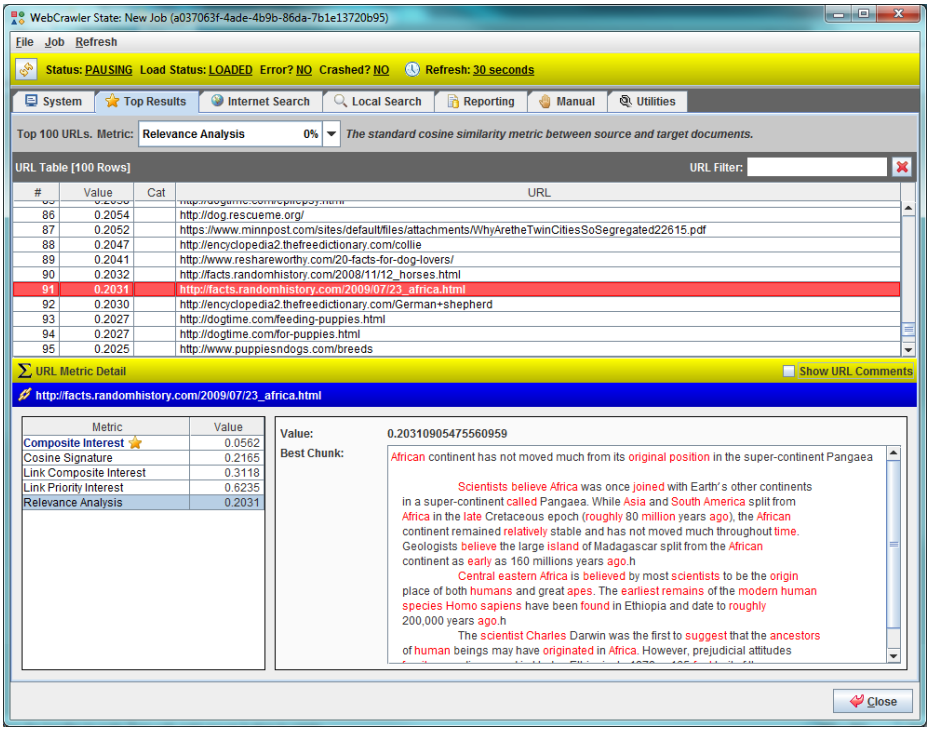

Figure 1: The Huntsman Analyst User Interface

All of these parameters are passed to Huntsman to begin the crawl. Huntsman submits keywords and phrases to various search engines to seed the crawler with good starting places on the internet. As Huntsman is crawling, each downloaded page is compared against the target signatures using a suite of text analytics. Throughout the crawl, the analyst is able to view the most relevant findings through the graphical user interface (see Figure 1). Huntsman also provides document excerpts and other explanations to the analyst regarding its reasoning for presenting this information to the analyst, allowing the analyst to make quick decisions about the importance of the information, as well as redirect the crawl as necessary. This interaction between the analyst and Huntsman continues until the analyst decides the quality of the data collection is sufficient.

After the crawl, the analyst is able continue to review and annotate the results, and is able to automatically generate a report with the most relevant findings and annotations for documentation or to present to others. This process can be seen in Figure 2.

Figure 2: The Intelligent Web Crawling Process

2.3 Huntsman as a Personal Research Assistant

In a sense, Huntsman can be viewed as a personal research assistant. This approach provides several distinct advantages; 1) analysts are able to perform a search that is targeted on the entire content of the documents, rather than just the presence of a few keywords; 2) Huntsman allows analysts to perform a more nuanced analysis of the document contents by applying a suite of text analytics and presenting the results to the analyst, as well as easy to understand explanations for why each document was considered interesting; 3) While Huntsman leverages the results of search engines, it moves beyond what search engines provide by analyzing all pages crawled and providing a rollup of the best results to the analyst; 4) Huntsman's search focuses on the content, not a search engine's assessment of the page's potential interest to the masses or to the individual; 5) Huntsman can peruse enormous quantities of information, saving the analyst time and allowing the analyst to better remain in context by providing focused results and reasoning behind those results.

2.4 Huntsman is Applicable to Various Contexts

Huntsman can be applied to multiple contexts such as data science, social modeling, business analysis, and field operations – essentially any situation that utilizes large datasets to make rapid, high-consequence decisions.

3 Conclusion

There are many benefits of using web crawling and text analytics in the analysis of large datasets. A web crawler is able to locate non-indexed information and does not rely on a search engine to serve as a middleman for compiling web data in a single dataset. Instead it uses search engines as a starting point and then crawls out from there to find any relevant data available on the open web.

Overall, there is a need to accurately and efficiently synthesize large amounts of information to enable decision-making. Huntsman is a versatile capability that has been developed and used across various contexts to assess large amounts of interest-

ing information, which ultimately reduces analysts' cognitive burden and applies findings to increase mission effectiveness.

References

1. Chakrabarti, S., Berg, M. v. d., & Dom, B. (1999). Focused crawling: A new approach to topic-specific Web resource discovery. *Computer Networks* 31 (11), 1623-1640
2. Henzinger, M., Heydon, A., Mitzenmacher, M., and Najork, M. (2000). On near-uniform URL sampling. *In Proceedings of the 9th International World Wide Web Conference*, pages 295–308, Amsterdam, Netherlands, May 2000. Elsevier Science.
3. Jasra, M. (2010). Google Has Indexed Only 0.004% of All Data on the Internet. http://www.webanalyticsworld.net/2010/11/google-indexes-only-0004-of-all-data-on.html
4. Zeinalipour-Yazti, D., & Dikaiakos, M. (2002)
5. Najork, M., & Wiener, J. L. (2001). Breadth-First Search Crawling Yields High-Quality Pages. WWW10, May 1-5, 2001, Hong Kong.
6. Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychological review*, *109*(1), 75.
7. Pope, C., Ziebland, S., & Mays, N. (2000). Analysing qualitative data. *Bmj*, *320*(7227), 114-116.
8. Howard, N., Spielholz, P., Bao, S., Silverstein, B., & Fan, Z. J. (2009). Reliability of an observational tool to assess the organization of work. *International Journal of Industrial Ergonomics*, *39*(1), 260-266.
9. Marchionini, G. (1997). *Information seeking in electronic environments* (No. 9). Cambridge university press.