IDATA Overview

Michelle Albert, Arun Maiya, Laura Odell, and Miranda Seitz-McLeese

The Challenge: Document review and analysis is typically a manual effort that requires significant amounts of staff time, and the sheer volume of documents makes it inevitable that manual reviewers will miss relevant, critical information. Instead of searching through documents, government personnel need to focus on the analysis and critical thinking necessary for informed decision-making.

IDATA moves beyond simple keyword search tools with its use of analytics-powered facets. ... Also, IDATA can search for numerical parameter values, a capability not available in common search engines.

IDATA Background

The Department of Defense (DoD) and other government agencies are increasingly faced with the challenge of analyzing large, heterogeneous data files in different formats (such as text, image, audio, and video) that cover a wide range of content. The ability to intelligently automate the analysis of big data sets to find relevant information is needed to solve problems quickly, accurately, and without bias. But there is no standard way to determine the information needed to solve a specific problem. This lack of a defined data collection process, as well as a reviewer's unavoidable bias (i.e., regarding the knowledge of the data's content, location, and relevance), has hampered the ability of government offices to respond quickly and thoughtfully to pressing issues. Many of the technology solutions currently available use simple word searches to identify relevant documents and data, but these are not sufficient.

IDA has made innovative uses of advances in text and discovery analytics by using supervised machine learning[1] and natural language processing[2] techniques to analyze the massive amount of documents and data available across DoD and the Federal Government. This capability, originally invented by Dr. Arun Maiya and called IDA Text Analytics (ITA)[3], has minimized the need for manual and repetitive human-intensive data collection processes and allowed more time for critical thinking and decision-making.

---

[1] Supervised machine learning is so named because the data scientist teaches the algorithm to arrive at the appropriate conclusions. Supervised machine learning requires the algorithm's possible outputs to be known and the data used to be labeled with the correct answers (Castle 2017).

[2] Natural language processing is a means for computers to analyze and derive meaning from human language. It focuses on the interaction between humans and computers, and combines computer science, artificial intelligence, and computational linguistics (Algorithmia 2017).

[3] IDA Text Analytics (ITA) was renamed IDA Text Analytics (IDATA) in 2018.

IDA's capability moves beyond traditional searching within information sets to a discovery approach that links information in different ways and builds on information triage principles. By automating the information triage process, we can rapidly collect and ingest documents and other types of data, discover previously unknown information, and support exploratory analysis to provide unique insights based on domain content. IDATA uses open source, state-of-the-art software and can be rapidly customized to address specific problems.

Combining IDA's information triage process with policy and operations expertise delivers information relevant to and necessary for solving individual problems while taking sponsors' business environments into account. This triage process has been used to examine internal memoranda and extract data to determine compliance, track changes in legislative language, determine policy and procedure alignment, and identify data transmission between information technology (IT) systems. As a result, sponsors have spent significantly less time finding and manipulating data, and more time analyzing relevant information.

Information Triage and IDATA's Process

IDATA combines peer-reviewed information triage principles with text analytics and exploratory data analysis techniques into a capability for gathering, sorting, and prioritizing information. It identifies what is relevant or important to the decision maker and discards everything else. The result is a richer overview of the entire information space, changing the focus from finding what the analyst knows exists to discovering information the analyst was not aware existed. IDATA also uses post-conditioning information processes based on text analytics, natural language processing, and supervised machine learning to facilitate exploratory data analysis.

IDA's information triage process helps refine the information gathered to focus on the most valuable information and identify additional information sources that may be relevant. It has three phases: the pre-conditioning phase, which involves data collection and pre-analysis; the interface phase, which involves search and discovery; and the post-processing phase, which involves post-analysis and data exploration. Figure 1 depicts the three phases.

In the pre-conditioning phase, subject matter experts identify publicly available data sources that could be relevant to the problem (e.g., Federal Register or Office of Management and Budget memoranda) and enter them into the IDATA capability. This creates a growing information repository that supports a range of content areas. In some cases, other DoD data sources, such as internal memoranda or reports, may also be collected to form a domain-specific document repository enclave. IDATA accepts data from a variety of digital sources, ranging from scanned hardcopy files to Excel spreadsheets to html/xml files. We developed a rich set of open-source tools that support converting these files into machine-readable formats.

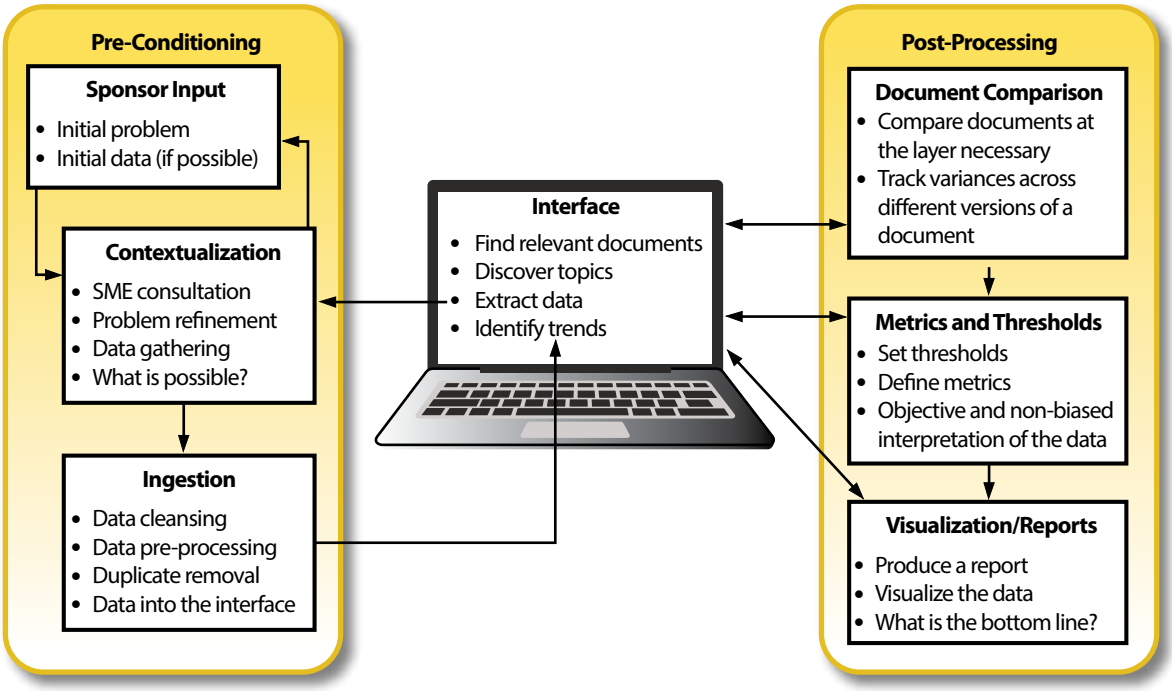
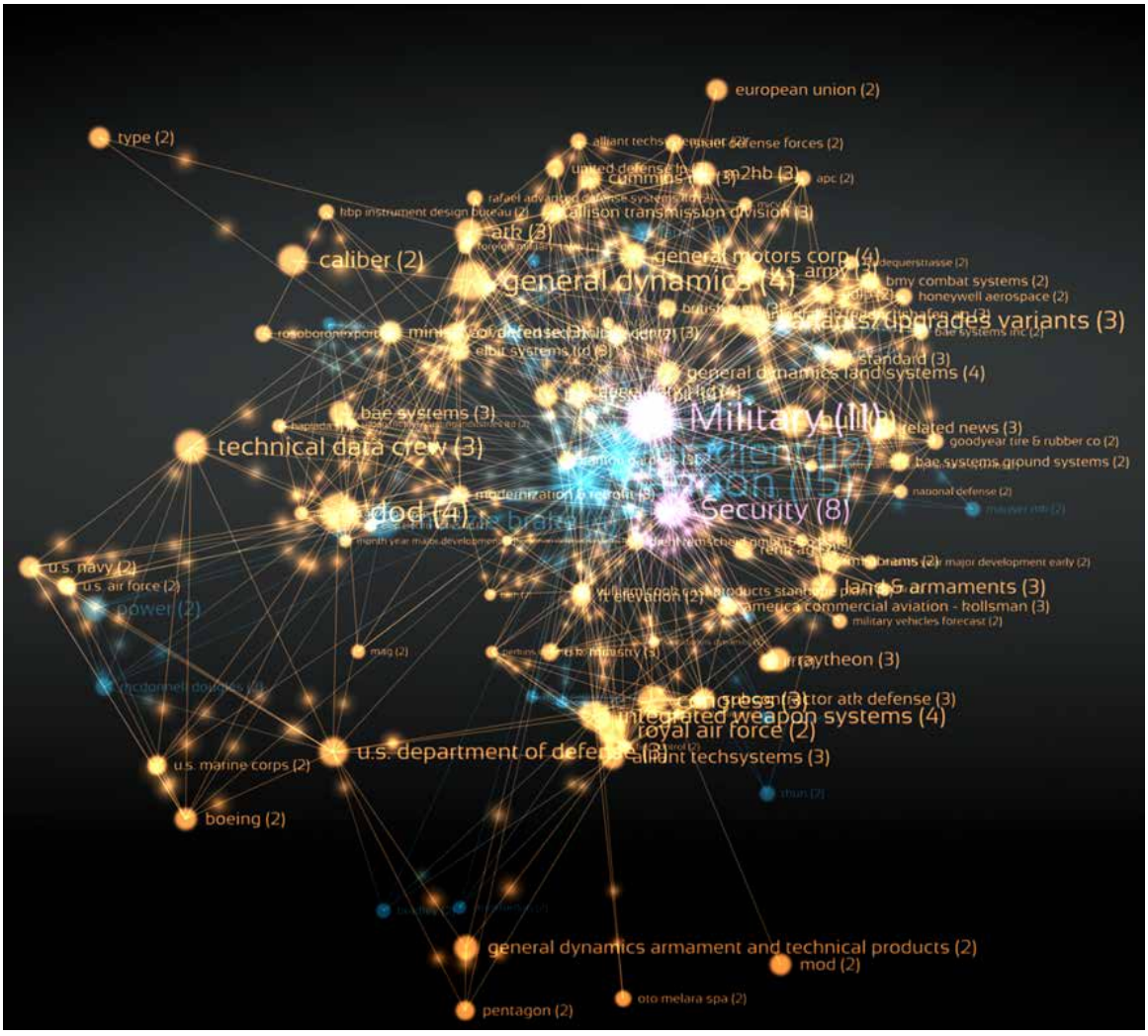


Figure 1. IDA Supervised Machine Learning Capability Process

In the second phase, documents in the repository are tagged or sorted in an efficient, automated fashion to extract information that will facilitate discovery. Documents can be selected through simple word searches or clustered into similar groups based on tags or data extraction categories. We created an interactive, customizable user interface that allows analysts to search the document repository for relevant information.

The interface allows subject matter experts to calibrate the criteria for finding certain documents. The discovery process begins with categorization and grouping as the analyst gradually filters out irrelevant data and focuses on critical information that the decision maker can use. What begins as an automated search problem (i.e., find all documents containing the phrase "machine learning") becomes an interactive discovery process (i.e., find all documents that contain information relevant to "machine learning").

In the post-processing phase, data exploration begins once IDATA finds a set of relevant documents. IDATA's set of automated machine learning and exploratory analysis tools create the capability to parse documents into smaller segments that can be compared across a document set, extract specific data from documents based on user-defined expressions, classify document segments with respect to specific problem domains, and provide a visual representation of the content in the document set (see Figure 2 for an example). These post-processing tools can be rapidly customized to suit government-specific problems, providing users with deep domain knowledge and the ability to extract pertinent information for analysis.



Note: This is a screen shot of an interactive graphic, displayed to indicate the complexity.

Figure 2. Visualization of an IDATA Search on Cybersecurity

What Is IDATA?

IDATA is a customizable software capability for exploratory analysis and triage of highly heterogeneous collections of documents (i.e., exploratory search). It is built on proven open-source components and uses different techniques based on machine learning and natural language processing to facilitate rapid insight discovery. IDATA supports both search (looking for specific information) and discovery (finding relevant information through interactive browsing) functions.

IDATA moves beyond simple keyword search tools with its use of analytics-powered facets (or filters) that allow a document set to be viewed along different dimensions or through various lenses. These facets, as well as other visualizations and auto-generated reports, provide rich overviews of the entire information

space and can use document research to answer questions. Also, IDATA can search for numerical parameter values, a capability not available in common search engines. IDATA uses multiple techniques to implement these facets, including (but not limited to):

- **Key phrase and concept discovery:** IDATA implements a key-term-extraction algorithm and displays an informative tag cloud of the most common discovered terms or concepts in a text data set. The tag cloud is based on computational linguistic techniques and enables the user to quickly gain a sense of the contents of any document set at a glance.
- **Topic clusters:** IDATA implements an unsupervised machine-learning-based topic modeling algorithm that automatically divides the documents in a data set into clusters. The documents in each cluster are semantically related to each other and to a general topic or theme. IDATA displays these data clusters as a menu showing document distribution across different themes. The user can click on a topic cluster to focus only on documents pertaining to the selected topic.
- **Supervised machine learning facets – technology area and document type:** IDATA can be trained to automatically group documents according to predefined categories by feeding it example documents representative of each category. In one configuration, we trained IDATA to categorize documents according to technology area (e.g., aeronautics; directed energy; lasers, optics, and sensors; positioning, navigation, and timing; signature control), using the Militarily Critical Technologies List as a guide. IDATA has also been trained to recognize documents based on report type (e.g., technical information, test plan, programmatic information).
- **Customizable entity extractions:** IDATA can discover and extract entities of interest (e.g., persons, organizations, locations) in a document set. The types of entities can be customized based on the needs of a particular application domain (e.g., military technical reports, tweets). For example, IDATA can identify measured quantities and reveal documents with sensitivity markings, such as "For Official Use Only," "FOUO," and "Law Enforcement Sensitive."
- **File metadata facets – location, time, and format:** IDATA supports filtering documents according to their location in the file hierarchy, which allows, for example, an analyst to view all files in the vicinity of a document that the analyst has determined to be of special interest. IDATA also supports filtering documents according to format (e.g., pdf, txt, docx, pptx, xslx) and the time of their last modification. Facets such as date and time can be customized to fit a particular need (e.g., date of publication, date of last modification).
- **Other features:** We continuously add new functions to the IDATA

capability, such as graph-based visualizations of text quantities, a means of detecting duplicates, and others. Visualizing query results using a semantic network can help the user get a holistic view of his or her search. Semantic networks allow for the abstraction of conceptual relationships. Graphing relationships between entities can grant insight into the overall structure of a system or significant statistical outliers. Figure 2 depicts an example of a 3D semantic network graph created with the output from an IDATA search.

IDATA Data Repository

IDA created a repository of publicly available unclassified federal policy documents informed by IDA's work for the Office of the Secretary of Defense. These documents include, but are not limited to:

- Reports from the House and Senate Armed Services Committees dating from 1991
- National Defense Authorization Acts (NDAA) dating from 1962
- United States Code in its entirety
- Federal Register dating from 2000
- Office of Management and Budget (OMB) bulletins, circulars, memoranda, and other documents
- Office of Information and Regulatory Affairs (OIRA) publications
- National Institute of Standards and Technology (NIST) publications
- National Archives publications
- DoD issuances
- DoD Financial Management Regulations
- Joint Chiefs of Staff issuances
- Federal Acquisition Regulation (FAR)
- Defense Federal Acquisition Regulation Supplement (DFARS)
- Defense Information Systems Network (DISN) memos
- Defense Information Systems Agency (DISA) memos
- DoD budget documents for the past five years
- Other DoD publications and documents

This repository serves as a starting point for IDATA searches. Additional, focused data collection is almost always required for each project. Each project's data collection adds to the overall repository and provides a rich foundation for follow-on research.

Documents are acquired in multiple ways. Sponsors have provided some data sets from their own collections. In a few cases, such as with collecting United States Code and Federal Register documents, we downloaded available bulk data zip files. Most often, we write a web crawler to look for relevant documents, automating as much of the data collection process as possible. Storing the data in a common database and reusing the repository for analysis reduce

the time spent searching for and aggregating documents.

Benefits of IDATA and Information Triage

IDA's information triage approach gives sponsors the opportunity to reduce costs in several ways, most notably by minimizing the time required to find, search, and analyze information across a variety of documents and file types. It can also find and process existing information, which eliminates duplicative activities stemming from an organization's inability to find or manipulate results from previous efforts.

The process of document review and analysis is still a mostly manual effort that requires significant amounts of staff time. The sheer volume of documents produced by government offices makes it inevitable that manual reviewers will miss relevant information. IDA's automated information triage approach removes much of the manual work and makes relevant, critical information easily available. Rather than spend a majority of time searching through documents, government personnel can instead focus on the analysis and critical thinking necessary for informed decision-making.

References

Algorithmia. "Introduction to Natural Language Processing (NLP) 2016." August 11, 2016. https://blog.algorithmia.com/introduction-natural-language-processing-nlp/. Accessed September 28, 2017.

Castle, N. "Supervised v. Unsupervised Machine Learning." *Datascience.com*. July 13, 2017. https://www.datascience.com/blog/supervised-and-unsupervised-machine-learning-algorithms. Accessed September 18, 2017.