IDA | RESEARCH NOTES

IDA Text Analytics

October 2018

IDA is the Institute for Defense Analyses, a non-profit corporation operating in the public interest.

IDA's three federally funded research and development centers provide objective analyses of national security issues and related national challenges, particularly those requiring extraordinary scientific, technical, and analytic expertise.

The articles in this edition of *IDA Research Notes* were written under the auspices of IDA's Information Technology and Systems Division (ITSD) and System Evaluation Division (SED). Please contact the ITSD Director, Dr. Margaret E. Myers (703.578.2782, mmyers@ida.org), or SED Director, Dr. Steve Warner (703.845.2096, swarner@ida.org), for more information on any of these articles.

IDA

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311
ida.org

The Institute for Defense Analyses has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Text Analytics

The articles in this issue of *IDA Research Notes* describe our research related to the development and application of the IDA Text Analytics (IDATA) capability.

IDATA development was originally motivated by an important Department of Defense (DoD) challenge: how to rapidly assess damage from cyberattacks in which engineering documents have been exfiltrated from defense industry networks. To address this challenge, IDA researchers developed the initial IDA Text Analytics capability, originally known as ITA. It is still in use today in the DoD Cyber Crime Center.

The opening article by ***Michelle Albert***, ***Arun Maiya***, ***Laura Odell***, and ***Miranda Seitz-McLeese*** describes IDATA itself and how it applies computational approaches in text mining and natural language processing to the discovery of critical information in unstructured collections of text.

In "Data Exploration and Management of Defense Finance and Accounting Services Artifacts," ***Laura Odell***, ***Robert Rolfe***, ***Andrew Wan***, and ***Anna Vasilyeva*** explain how they piloted use of the IDATA capability to discover useful data in DoD business information systems, which are spread across fragmented, unstructured, and inconsistent sources. The approach described in the article helped make valuable information available for analysis and demonstrated that the IDATA capability has the potential for application to other data sets to make valuable information available for analysis.

"Extracting Structured Numerical Data from Large Quantities of Memoranda" describes how ***Laura Odell***, ***Miranda Seitz-McLeese***, and ***James O'Connor*** used the IDATA capability to help the Defense Logistics Agency understand the cumulative variances of actual and expected weights of natural resource stockpile materials without having to reweigh the stockpiles, which was not feasible.

Next, ***Laura Odell***, ***Kathy Burton***, and ***Miranda Seitz-McLeese*** in the article "Implementing the Federal Advisory Committee Act" discuss how they used the IDATA capability to provide a timely, comprehensive analysis of DoD Federal Advisory Committee Act (FACA) processes. The IDA researchers' work resulted in a change to DoD policy regarding FACA and associated procedures for vetting and appointing members to DoD's advisory committees. The researchers performed this analysis in less than a week; without the IDATA capability, it would have taken significantly more time (months, at least) to manually collect relevant documents and identify sections in those documents that were pertinent to the questions posed.

Building on the FACA project, the same three researchers applied the IDATA capability to tackle the challenge of developing a list of all recurring reports that DoD is responsible for submitting to Congress ("Finding and Categorizing Recurring Reports to Congress") and then to develop a process for tracking changes between the House and Senate versions of a National Defense Authorization

Act ("Comparing the House and Senate Versions of the National Defense Authorization Act"). Using IDATA to streamline a manual, repetitive, and time-consuming process significantly shortened the research phase, leaving more time for analysis.

"Discovering, Analyzing, and Understanding Improvised Explosive Device Documents" reports on research conducted by ***Forrest Frank*** using IDATA to help the Joint Improvised Explosive Device Defeat Organization (now the Joint Improvised-Threat Defeat Organization) improve its understanding of science and technology projects from across the government aimed at defeating improvised explosive devices.

"Use of IDATA Capabilities for Social Media Analytics" discusses how ***Thi Uyen Tran*** and ***Daniel Nakada*** used the IDATA capability to help the Defense Threat Reduction Agency understand how well a biosurveillance application could find relevant disease information on social media.

The concluding article by ***Michelle Albert*** provides background information on the award-winning IDATA team and discusses future opportunities for applying IDATA to solve real-world problems.

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.

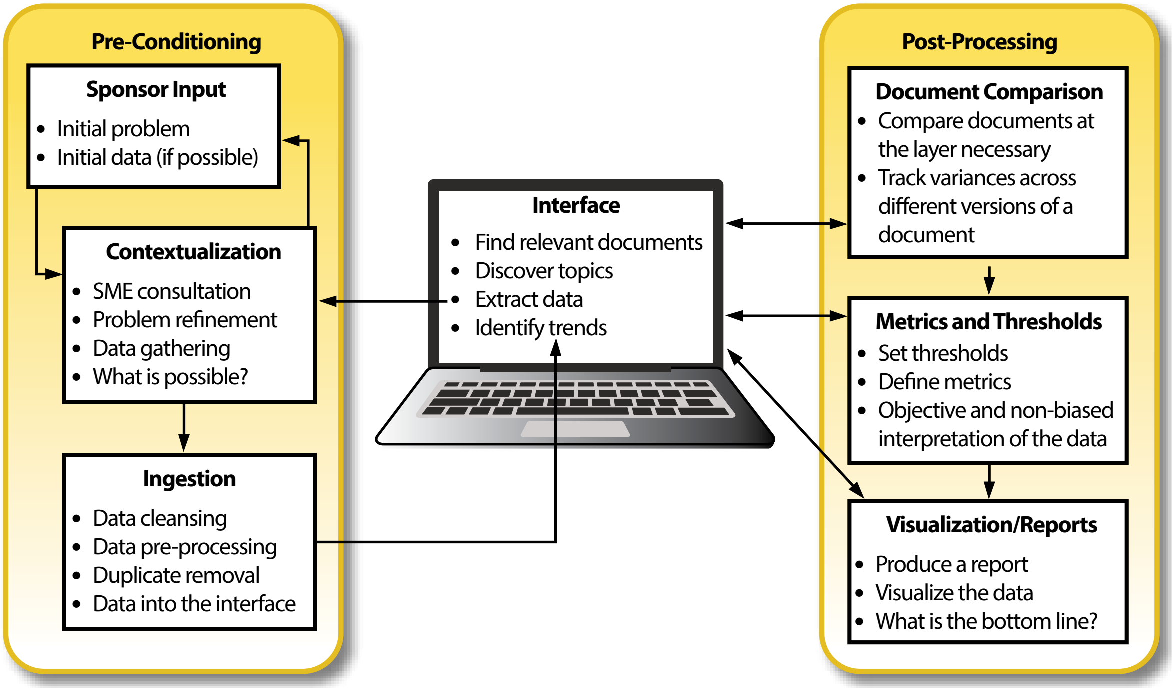
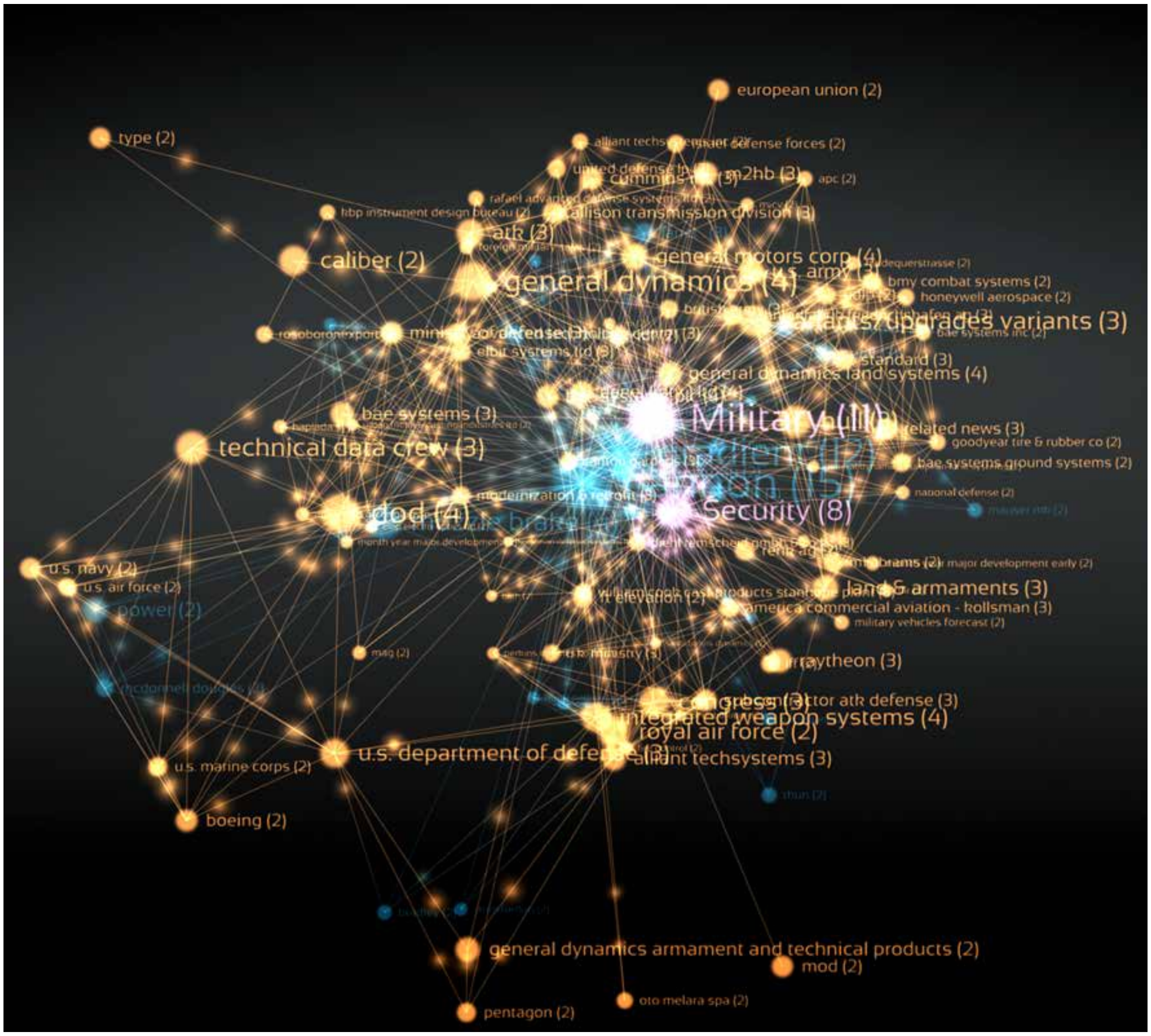


Figure 1. IDA Supervised Machine Learning Capability Process

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.

Data Exploration and Management of Defense Finance and Accounting Services Artifacts

Laura Odell, Robert Rolfe, Andrew Wan, and Anna Vasilyeva

The Challenge: DoD's business information systems contain useful data, but they are spread across fragmented, unstructured, and inconsistent sources. The Department needed better methods to make valuable information available for analysis.

No single, comprehensible data source provides the information needed to model and understand DoD's business system network. The information is scattered...

Background

The information systems that support DoD business processes comprise a vast and complex network of interactions related to data collection, transmission, and summation. These systems support activities ranging from accounting and procurement, to payroll, to travel. To improve efficiency, reduce costs, and determine the effects of impending changes, decision makers need to be able to reliably explore and analyze these systems and their interactions.

However, no single, comprehensible data source provides the information needed to model and understand DoD's business system network. The information is scattered throughout the unstructured text of roughly 1,000 memoranda and interface control documents in several structured, but incomplete, repositories.

DoD Chief Information Officer (CIO) Business Process Reengineering (BPR), in collaboration with the Office of the Under Secretary of Defense for Acquisition, Technology and Logistics (OUSD(AT&L)) asked IDA to analyze a Defense Financing and Accounting Services (DFAS) data set. This data set was one of the many used as training sets for the project.

Methodology

IDA researchers' work on the DFAS data set comprised three separate but concurrent efforts: extracting structured information from unstructured text; combining structured and unstructured data sets that present conflicting views of the data; and developing ways to navigate, search, analyze, validate, and correct the resulting information.

IDA extracted and combined data from thousands of DFAS agreements, DoD Information Technology Portfolio Repository (DITPR) entries, line items from the DoD Information Technology

Budget Estimates to Congress, and DFAS 7900.4-M, *Financial Management Systems Requirements Manual*. We then categorized the extracted data into three types: entities, relations, and entity attributes. For this task, a relation refers to both the entities that entered into an agreement and the data sharing between entities. We found 422 entities comprising information systems, organizations (that operate or own a system), modules, and other types that participate in agreements. IDA also collected information about each entity, including budget size, business function, and a description from DITPR.

IDA then created a knowledge base about each system and its interactions. We adapted natural language processing (Bird, Klein, and Loper 2009) and machine-learning techniques (Flach 2012) to automate the initial data extraction and aggregation, which would have been unmanageable if approached manually. The systems and their interactions made up a network of more than 400 nodes and 1,000 edges.

Figure 1 illustrates the process through four interdependent activities: information extraction, data merging, data processing, and exploration and analysis. The percentages in the lower right-hand corner of the blue boxes estimate the amount of work that could be automated.

The process diagram shows the existing knowledge base as a controlling factor in the extraction output, which reflects IDA's finding that the ability to extract information is influenced by the amount and quality of structured data that already exist. The diagram also suggests that

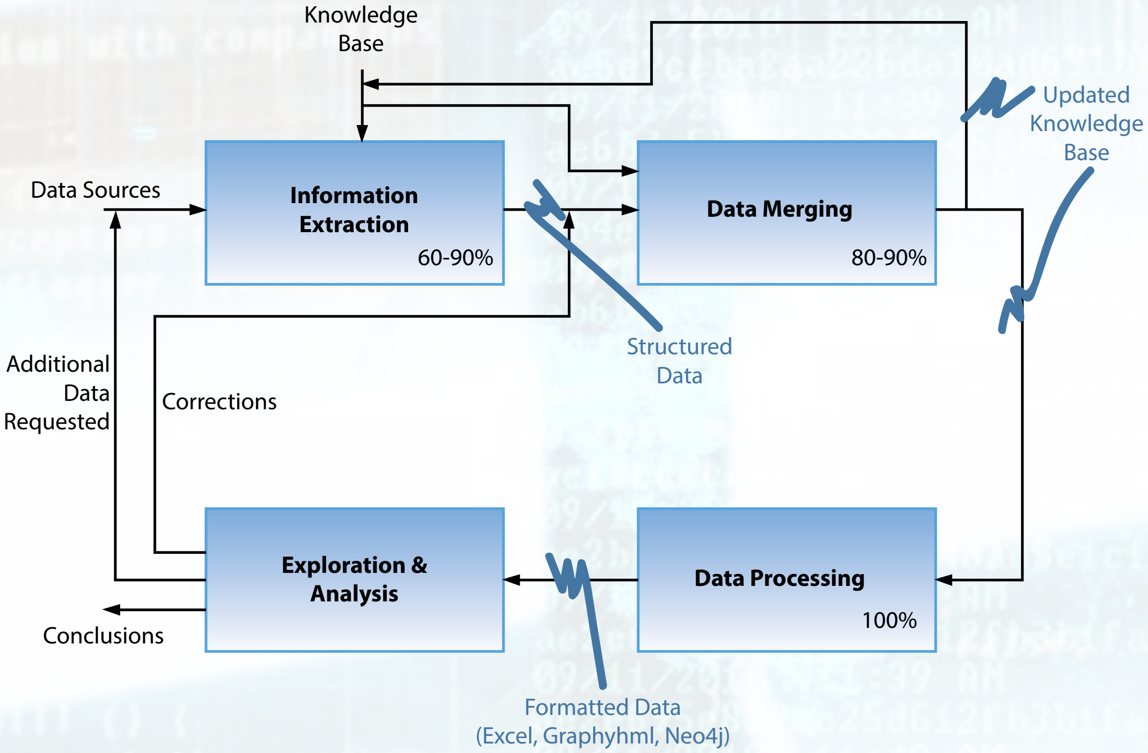


Figure 1. Process Diagram

both the types of sources considered and the quality of the extraction will depend on the results of other activities downstream; for example, corrections that result from exploration and analysis will affect the existing knowledge base, and thus extraction.

We determined the most time-consuming activity to be information extraction – the process of obtaining structured data from unstructured data sources. Information extraction can be divided into several sub-activities with complex dependencies: strategy design for extraction based on project goals and the properties of data sources, pre-processing to transform various data formats, entity and relation extraction (Freitag 2000), and manual intervention. IDA's success in automating the information extraction process varied among sub-activities. Entity extraction was fairly accurate and fast, but relation extraction was less successful.

Results

This effort resulted in a knowledge base of detailed information about each system and how it interacts with other systems in the DFAS network. IDA's adaptation of techniques from natural language processing and machine learning to automate the initial extraction and aggregation made the manual refinement of an otherwise unmanageable, complex array of information possible. We merged the resulting information with other data sources to add detail, again using a combination of automated and manual efforts.

To enable further exploration and analysis, we used an open source software platform originally developed for visualizing and analyzing biomolecular networks (Cytoscape n.d.) to display the data in a graph (Figure 2). Multiple system, edge, and network attributes[1] can control the graph's appearance and be used to navigate the data through user-defined filter and search queries.

The nodes represent applications and offices. The connections between nodes depict data flow through memorandums of agreement (MOA), contracts, and other vehicles. The graph uses micro data to create a macro view. It shows how individual nodes and groups of nodes are connected. These connections provide insight into what might be affected if a node (or group of nodes) or a particular data flow changes.

IDA was able to use these methods to automatically produce useful data from imperfect sources. Because a rigorous analysis requires validation and correction, the researchers also provided means of quickly accessing supporting documents and information while exploring the data in the graph. The data can then be updated and a new visualization generated.

Impact

DoD's existing assets contain useful data, but they are hidden in fragmented, unstructured, and inconsistent sources. The network of information systems that support DoD

[1] Attributes are the structured data produced by the natural language processing and machine learning processing of the raw data.

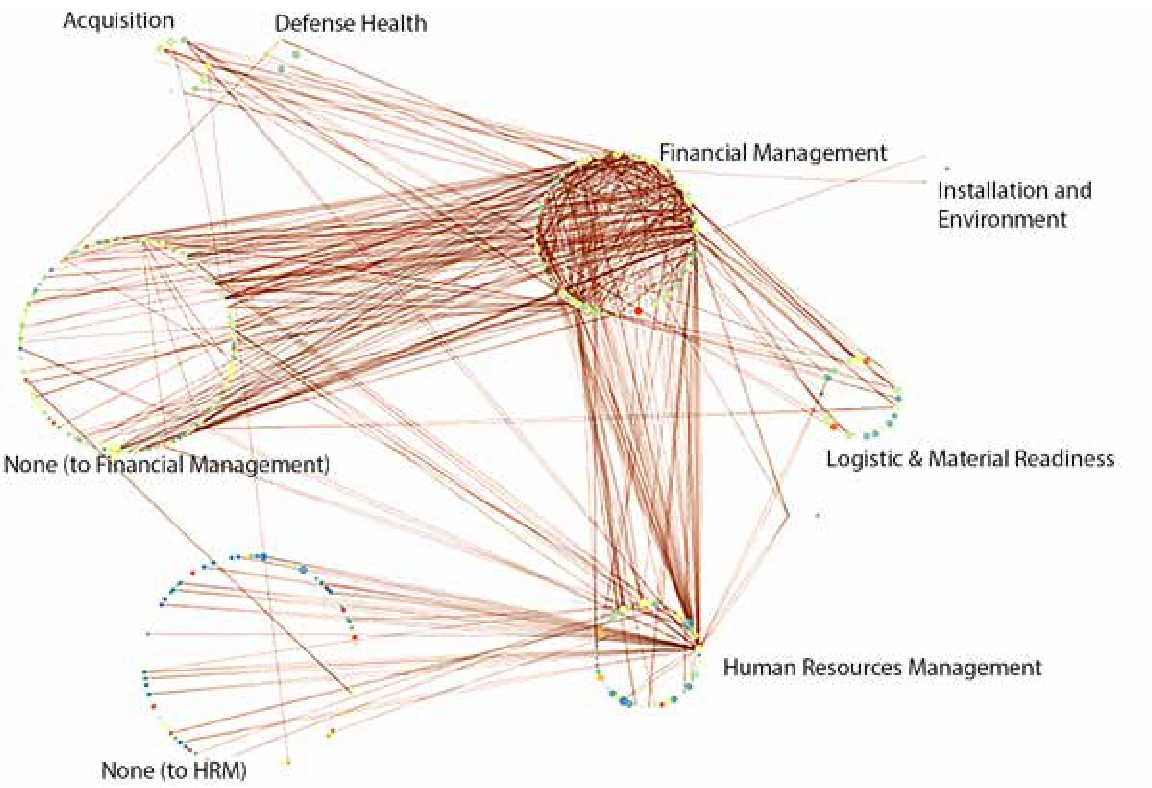

Acquisition
Defense Health
Financial Management
Installation and Environment
None (to Financial Management)
Logistic & Material Readiness
Human Resources Management
None (to HRM)

Figure 2. Group Attributes Layout by Function

business and its latent presentation[2] in DFAS agreements is just one example of this pervasive phenomenon.

The methods that IDA developed can be applied to other data sets to make valuable information available for analysis. The methods made what otherwise would have been a monumental task feasible.

References

Bird, S., E. Klein, and E. Loper. 2009. *Natural Language Processing with Python*. Beijing: O'Reilly.

Cytoscape. n.d. http://www.cytoscape.org.

Flach, P. 2012. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. New York, NY: Cambridge University Press.

Freitag, D. 2000. *Machine Learning for Information Extraction in Informal Domains*. Dordrecht: Kluwer Academic Publishers. http://www.cs.bilkent.edu.tr/~guvenir/courses/CS550/Seminar/freitag2000-ml.pdf.

[2] *Latent presentation* is a mathematical term that refers to making assumptions about a large data set using only available data (that is, some data are not available or accessible). In this case, DoD has a known network of information systems, but it is not feasible to observe each system and connection in the network due to its size. Instead, it was feasible to gather MOAs and other agreement documents and use them to create a partial map of the network. We could then use this map to make assumptions about the entire DoD network.

Extracting Structured Numerical Data from Large Quantities of Memoranda

Laura Odell, Miranda Seitz-McLeese, and James O'Conner

The Challenge: The Defense Logistics Agency needed to understand the cumulative variances of actual and expected weights of natural resources stockpile materials. Physically reweighing the stockpiles to determine the differences between the expected and recorded amounts was not feasible.

Background

The Defense Logistics Agency (DLA) is tasked by executive order with managing the nation's stockpile of strategic materials. As part of DLA's 2015 audit-readiness effort, it needed to understand the cumulative variances of actual and expected weights of natural resources stockpile materials. Since the data were not readily available, auditors suggested that the DLA Strategic Materials Division (SMD), which oversees the strategic materials stockpile, reweigh the piles to determine the differences between the expected and recorded amounts. DLA SMD determined that reweighing the piles was not practical because reweighing would mean physically moving material, which is expensive, time-consuming, and labor-intensive, and could affect the environment. Also, for some materials, the reweighing process itself can cause material loss or degradation.

As part of DLA's 2015 audit-readiness effort, it needed to understand the cumulative variances of actual and expected weights of natural resources stockpile materials.

Instead, DLA turned to IDA for an alternative solution. Although DLA had not kept the data required to determine the difference between the expected and actual weights, it had maintained paper copies of 549 memoranda documenting the transaction or event details needed to assess whether material stockpiles were within a generally acceptable range of loss or gain when compared to industry benchmarks. IDA determined that, if we could extract numerical data from these memoranda, we could use those data to calculate whether the cumulative variances fell within industry standards.

Why Text Analytics?

Extracting the numerical data necessary to calculate variances from a stack of paper without a text analytics capability would have been labor-intensive. Someone would have had to read each memorandum, find the relevant numbers, and enter them into a spreadsheet. In addition, these data needed to be at an audit-ready level, which requires a small margin of error; the margin of error for manually entered data would be too high. Using IDATA allowed IDA researchers to produce high-quality data quickly.

The Process

DLA scanned the documents into JPEG files, but the scans had no text data associated with them, and IDATA cannot perform text analysis on documents that do not have text data. Our first step was to extract the text from the scanned documents using Adobe Acrobat's Optical Character Recognition (OCR) tool. Once the OCR tool extracted the text, we used IDATA's extractor tool to create structured data in the form of a spreadsheet. Extractors find information buried in text data according to certain patterns. For DLA's memoranda, IDA researchers wrote an extractor that pulled the dates of each memorandum based on their predictable structure (month, day, year). We wrote another extractor to find the reason given for any discrepancy based on the assumption that these reasons usually occurred in phrases such as, "…caused by [reason]."

Because DLA provided the entire data set, the search and discovery phase of IDA's information triage approach was not needed, and the data set moved to the exploratory analysis phase. IDA ran the text data through the extractors and entered the output into a spreadsheet. We reviewed the completed spreadsheet and adjusted the extractors to improve performance. We also had to input some data by hand because some of the documents were of such poor quality that the OCR tool could not read them. We also performed random spot checks to ensure the accuracy of the extractors and checked the completed spreadsheet for missing, incorrect, or duplicative entries, which we corrected manually. This process took two IDA researchers less than a week of work to complete.

Results

IDA used its information triage process to identify 469 instances of stockpile material measurements with associated reasons for weight differences from the 549 memoranda. We divided the causes into categories:

- Scale variance between measurements (i.e., equipment discrepancies)
- The environmental effect on the stockpile (i.e., snow, rain, type of ground)
- Administrative error (i.e., human error and equipment failure)
- Moisture evaporation over time
- Theft (this occurred only once, but we considered it important to include)
- Other (i.e., damage to equipment, no available explanation, comingling of piles).

Of the documents entered in the spreadsheet, none were missing the subject field, four were missing a cause statement, five were missing the shipment date, 10 were missing a dollar amount, two were missing a weight difference, 17 were missing the acquisition rate, and 78 were missing the percent over or under the original expected weight. We restricted further analysis to the 391 documents that had information about the variation in terms of the percentage of the original expected weight.

IDA used pivot tables and charts to determine the degree of variance between the expected weight and the actual weight and found that the variance was almost an order of magnitude lower than the industry standard. These results convinced DLA's auditors that reweighing the piles was unnecessary, which saved DLA significant time and money.

Implementing the Federal Advisory Committee Act

Laura Odell, Katharine Burton, and Miranda Seitz-McLeese

The Challenge: DoD officials needed to assess rapidly options for improving and streamlining DoD implementation of the Federal Advisory Committee Act while remaining compliant with federal policy and regulations.

DoD was particularly interested in identifying the differences between DoD-originated requirements in its FACA processes and policy and regulatory requirements from external agencies.

Background

The office of the DoD Chief Information Officer (CIO) Business Process and Systems Review (BPSR) requested IDA's assistance in answering two questions concerning DoD regulatory requirements for Federal Advisory Committees.

The first question involved assessing how proposed legislative changes to the Federal Advisory Committee Act (FACA) would affect DoD. FACA defines how federal advisory committees operate and requires open meetings, chartering, public engagement, and reporting (P.L. 92-463. Federal Advisory Committee Act (FACA) 1972). DoD was particularly interested in identifying the differences between DoD-originated provisions in its FACA processes and policy and regulatory requirements from external agencies.

The second question concerned analyzing stakeholder feedback on the Federal Advisory Committee management process. DoD asked IDA to identify common themes in the feedback and determine whether DoD was able to control or influence potential solutions.

The following discussion focuses on the analysis performed for the first question. Figure 1 illustrates the FACA policy hierarchy relevant to this task.

Methodology

For this project, IDA supplemented the IDATA existing document repository with documents from the Office of Government Ethics. We conducted a phased analysis of the information and began by identifying, collecting, and organizing the information that concerned the FACA. The IDATA capability facilitated information collection and analysis by identifying relevant documents and conducting a breakdown comparison of pertinent sections of the documents under investigation.

The search and discovery phase of IDA's information triage process began with a simple key word search to identify

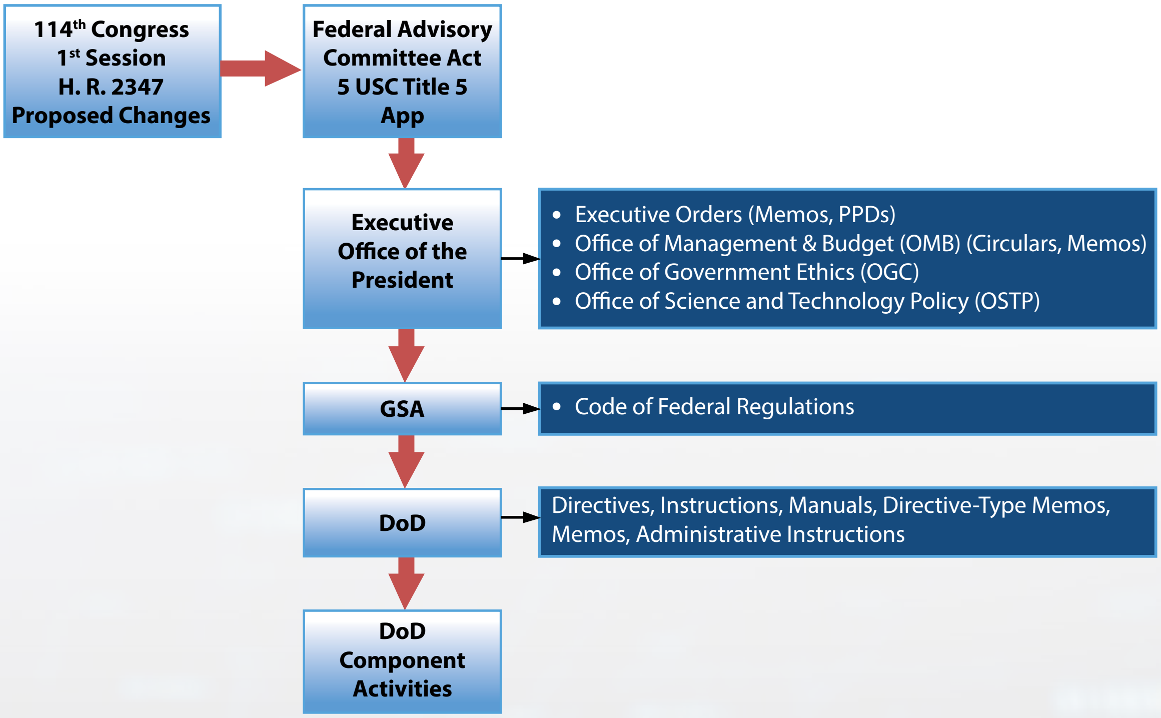

114th Congress
1st Session
H. R. 2347
Proposed Changes
Federal Advisory Committee Act
5 USC Title 5 App
Executive Office of the President
• Executive Orders (Memos, PPDs)
• Office of Management & Budget (OMB) (Circulars, Memos)
• Office of Government Ethics (OGC)
• Office of Science and Technology Policy (OSTP)
GSA
• Code of Federal Regulations
DoD
Directives, Instructions, Manuals, Directive-Type Memos, Memos, Administrative Instructions
DoD Component Activities

Figure 1. FACA Policy Hierarchy

regulatory and statutory documents from DoD, the General Services Administration (GSA), the Executive Office of the President, and Congress. The IDATA capability grouped these documents according to similarities in content and language. This allowed us to identify not only well-known DoD and federal policy and guidelines but also policy from smaller organizations that affected DoD's FACA policy. Of 500,000 publicly available documents associated with all federal and DoD policy, the IDATA capability identified one relevant DoD issuance and eight additional legislative and federal policy documents that affected DoD.

We converted the documents to XML to impose a hierarchical structure that allowed the documents to be segmented into relevant sections. We then inserted these sections into a machine-learning pipeline of processes and algorithms developed using the open-source library scikit-learn. IDATA removed conjunctions, articles, and pronouns ("if," "and," "the," and "it"), split the sections into words, and created word bigrams for each section, which were indexed using the term frequency-inverse document frequency (TF-IDF)[1] metric expressed in this equation:.

$$\left(1 + \mathit{log}\ (\#\ \mathit{appearances\ in\ document})\right) * \log\left(\frac{\mathit{total}\ \#\ \mathit{documents}}{\#\ \mathit{containing\ the\ term}}\right)$$

[1] TF-IDF weights a given term to determine how well the term describes an individual document within a corpus of documents. It does so by both weighting the term positively for the number of times it occurs within a specific document and weighting the term negatively relative to the number of documents that contain it (tfidf.com, http://www.tfidf.com/. Accessed September 26, 2017).

This process compared the sections and bigrams from the DoD issuance with the sections and bigrams from the other legislative and federal policy documents. The process yielded a matrix of TF-IDF values for each section-bigram combination. The researchers then used Latent Semantic Analysis (LSA)[2] to reduce the TF-IDF matrix to a smaller version containing all of the relevant sections but only the columns that captured the most variance between sections.

We used the smaller matrix to identify the most likely source for each section of the DoD issuance and tagged the sections to note the part of the issuance they came from. The analysis focused on the sections of the issuance that actively placed requirements on DoD. The researchers then used a threshold variance of 1.07 to determine whether a difference was present between sections.

Our analysis answered four questions:

1. *What is the source of DoD issuance requirements?* Using the matrix that resulted from the LSA, we identified the most likely source for each section of the DoD issuance.

2. *What is the crosswalk from statute to DoD issuance?* For completeness, the DoD issuance was compared with all of the documents to determine how requirements flow from Congress to DoD. We applied an agglomerative centering method to the matrix from the LSA to trace the requirements from the DoD issuance across the FACA policy landscape. The algorithm begins with the issuance and works its way outward, from more general documents to more specific documents.

3. *What is the potential impact of proposed legislation on the current statute?* We aligned sections of proposed legislation with the current legislation to reveal not only changes in language but also the locations of the language in the original and proposed statutes.

4. *What DoD issuances mention FACA and may be affected by any changes to the instruction?* We used the search capability in the IDATA document repository to identify all DoD issuances that cited FACA.

Similarities between sections suggest requirements imposed on DoD by legislation or other federal policy; differences between sections suggest DoD-imposed requirements. We found that the differences were primarily in the procedures sections of the documents.

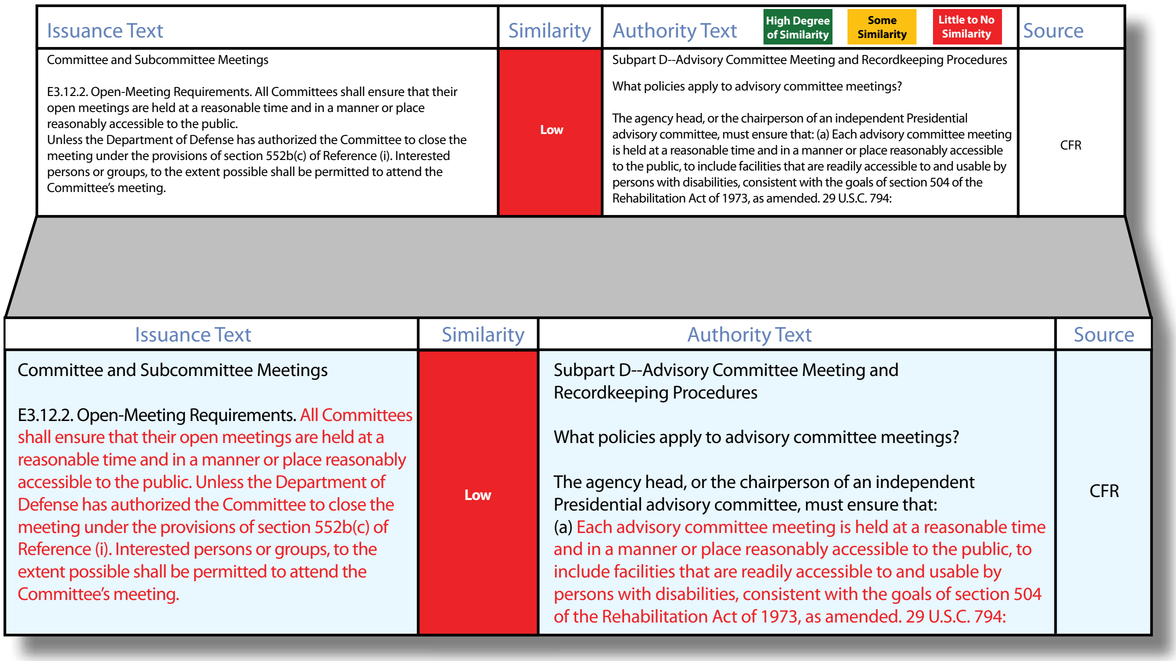
Results and Impact

The algorithm ranked the sections according to three criteria: (1) the raw number of sections that registered as "significantly different"

[2] LSA is a method for determining the similarity in the meaning of words and phrases by analyzing a large corpus of text and producing a set of related concepts and terms. LSA is known to combat the effects of synonymy (a state in which a word is a synonym for other words) and polysemy (that a word or phrase may have more than one meaning).

from text in other guiding documents, (2) the percentage of sections that registered as “significantly different,” and (3) the extremity of the difference. Based on these criteria, we were able to interpret the results, identify the eight documents that contained binding guidance from other agencies, and compare those documents with DoD’s procedures. The numerical results also helped us find the sections of the DoD issuance that were most likely self-imposed requirements. Figure 2 shows a sample of the results. In the figure, “Issuance Text” refers to the DoD document and “Authority Text” refers to the other legislative and federal policy documents. “CFR” in the figure is the Code of Federal Regulations.

IDA’s work resulted in a change to DoD policy regarding FACA and associated procedures for vetting and appointing members to DoD’s advisory committees. The researchers performed this analysis in less than a week; without the IDATA capability, it would have taken significantly more time (months, at least) to manually collect relevant documents and identify sections in those documents that were pertinent to the questions posed. The IDATA capability enabled a timely, comprehensive, and unbiased analysis that afforded DoD the time needed to evaluate opportunities to improve and streamline its FACA processes while remaining compliant with federal policy and regulations.



Issuance Text	Similarity	Authority Text 🟩 High Degree of Similarity 🟨 Some Similarity 🟥 Little to No Similarity	Source
Committee and Subcommittee Meetings E3.12.2. Open-Meeting Requirements. All Committees shall ensure that their open meetings are held at a reasonable time and in a manner or place reasonably accessible to the public. Unless the Department of Defense has authorized the Committee to close the meeting under the provisions of section 552b(c) of Reference (i). Interested persons or groups, to the extent possible shall be permitted to attend the Committee’s meeting.	🟥 Low	Subpart D--Advisory Committee Meeting and Recordkeeping Procedures What policies apply to advisory committee meetings? The agency head, or the chairperson of an independent Presidential advisory committee, must ensure that: (a) Each advisory committee meeting is held at a reasonable time and in a manner or place reasonably accessible to the public, to include facilities that are readily accessible to and usable by persons with disabilities, consistent with the goals of section 504 of the Rehabilitation Act of 1973, as amended. 29 U.S.C. 794:	CFR

Issuance Text	Similarity	Authority Text	Source
Committee and Subcommittee Meetings E3.12.2. Open-Meeting Requirements. All Committees shall ensure that their open meetings are held at a reasonable time and in a manner or place reasonably accessible to the public. Unless the Department of Defense has authorized the Committee to close the meeting under the provisions of section 552b(c) of Reference (i). Interested persons or groups, to the extent possible shall be permitted to attend the Committee’s meeting.	🟥 Low	Subpart D--Advisory Committee Meeting and Recordkeeping Procedures What policies apply to advisory committee meetings? The agency head, or the chairperson of an independent Presidential advisory committee, must ensure that: (a) Each advisory committee meeting is held at a reasonable time and in a manner or place reasonably accessible to the public, to include facilities that are readily accessible to and usable by persons with disabilities, consistent with the goals of section 504 of the Rehabilitation Act of 1973, as amended. 29 U.S.C. 794:	CFR

Figure 2. Sample Output of FACA Analysis

Reference

Federal Advisory Committee Act (FACA), Pub. L. 92-463, 86 Stat. 770. October 6, 1972. tfidf.com. http://www.tfidf.com/. Accessed September 26, 2017.

Finding and Categorizing Recurring Reports to Congress

Laura Odell, Katharine Burton, and Miranda Seitz-McLeese

The Challenge: DoD had no single source listing the recurring reports that DoD was responsible for submitting to Congress.

Background

DoD is required to send multiple reports to Congress, and it is difficult to keep track of them all – when they are due, what they must contain, which office receives which report. Attempting to manually track each report would require months of sustained work. Because National Defense Authorization Acts (NDAA) and other public laws add, modify, or remove reporting requirements, manually maintaining an updated list of reports would also require significant effort and time.

Because National Defense Authorization Acts (NDAA) and other public laws add, modify, or remove reporting requirements, manually maintaining an updated list of reports would also require significant effort and time.

The Office of the Under Secretary of Defense for Acquisition, Technology and Logistics (OUSD(AT&L)) asked IDA to identify sections of Title 10, U.S. Code that imposed recurring reporting requirements on DoD, as well as the frequency of the reporting requirements.

Methodology

The U.S. Code is available online in XML format. We split the XML-structured version of Title 10 into sections – several thousand sections at the start – and began to identify the sections that imposed reporting requirements. In machine learning, classification is a means of determining whether an object (in this case, a section of Title 10) belongs to a certain set (or group of related objects). Classification is a common machine learning task, but most classification algorithms require a training data set before they can be applied. This task did not have a training data set. And, because the sponsor had requested IDA to minimize manual effort, manually flagging several thousand documents as "imposing requirements" or "not imposing requirements" was not practical.

Instead, we used regular expressions[1] to find a small subset of documents that impose reporting requirements. Although this subset was too small to use as a training data set for a

[1] A regular expression is a special text string that describes a search pattern (RegularExpressions.info, http://www.regular-expressions.info/. Accessed September 26, 2017).

robust classification algorithm, it was large enough for the researchers to conduct a meaningful statistical analysis. We used Bayesian techniques to identify words and phrases that were statistically more likely to indicate a reporting requirement. The researchers then wrote a simple classification algorithm based on the results of the analysis.

Once we identified the sections that imposed reporting requirements, we began extracting metadata, including report frequency, subject, and responsible office. We sorted documents by extracting terms associated with frequency and periodicity (e.g., "annual," "quarter"). We then took a sample of the remaining documents and sorted them according to those extracted terms. These phrases were added to the extraction, and the process was repeated until only a few documents remained, which had to be manually sorted.

We used Title 10's chapter headings as descriptions of the subject matter of the required reports. We manually collected the headings to create a starter training data set, and used a label propagation algorithm to group the reports under general topic areas. We then extracted the office responsible for each report using a function designed for a previous project. The function uses regular expressions and string matching to identify agencies and offices under the Secretary of Defense. The researchers created a table listing the text, citation, subject, topic area, and periodicity of the reporting requirement from each section of Title 10 and submitted it to the sponsor.

Results

IDA found about 200 sections of Title 10 that imposed a recurring reporting requirement. The majority of these were annual reports, although biannual and biennial reports also figured prominently. Quarterly and quadrennial reports were the least frequent. Following annual reports, the most common type was event-triggered reports, or reports that required submission to Congress after a particular event occurred. Event-triggered reports are the type that DoD is most likely to lose track of, especially if the events triggering the report happen rarely.

Impact

Before IDA's analysis, no single source listed the recurring reports that DoD was responsible for submitting to Congress. IDATA's automation capabilities saved manpower and resources: the effort was conducted by a single analyst and a few subject matter experts over a few weeks. The process used to generate the report list could also easily be adapted to update an existing list.

DoD can use the report list to improve allocation of scarce resources to the reports that need immediate attention and be prepared to tackle event-triggered reports, avoiding surprises.

References

Glickman, M.E. and D.A. van Dyk,. "Basic Bayesian Methods," in W.T. Ambrosius, (Ed.), *Methods of Molecular Biology, Vol. 4: Topics in Biostatistics.* Totowa, NJ: Humana Press, Inc. 2007.

RegularExpressions.info. http://www.regularexpressions.info/. Accessed September 26, 2017.

Comparing the House and Senate Versions of the National Defense Authorization Act

Laura Odell, Katharine Burton, and Miranda Seitz-McLeese

The Challenge: The current process for tracking changes between the House and Senate versions of a National Defense Authorization Act is manual, repetitive, and time-consuming, leaving little time for analysis.

To save resources and allocate more time to analysis, the Office of Legislative Affairs asked IDA to see whether the IDATA capability could automate the text comparison process.

Background

The DoD Office of Legislative Affairs spends significant resources comparing different versions of the National Defense Authorization Act (NDAA) in a process that has not changed in two decades. The current process for tracking changes between the House and Senate versions of an NDAA is manual and repetitive. Analysts compile tables with the House language on one side and the Senate language on the other, and then examine the text for differences. Simply searching for differences in the text takes up so much time that the subsequent work of analyzing the potential impact of the discovered differences or determining which version is likely to be present in the final version become secondary priorities.

To save resources and allocate more time to analysis, the Office of Legislative Affairs asked IDA to see whether the IDATA capability could automate the text comparison process.

Methodology

Draft legislation is available online in XML format. IDA researchers downloaded the XML files and used the XML structure to split them into smaller sections. Researchers used these smaller sections to generate a term frequency-inverse document frequency (TF-IDF)[1] matrix, and then used latent semantic analysis (LSA)[2] to transform the matrix into a smaller dimensional vector space. Once the points were

[1] TF-IDF weights a given term to determine how well the term describes an individual document within a corpus of documents. It weights a term positively for the number of times the term occurs within a specific document and weights the same term negatively relative to the number of documents that contain it (tfidf.com, http://www.tfidf.com/Accessed September 26, 2017).

[2] LSA is a method for determining the similarity in meaning of words and phrases by analyzing a large corpus of text and producing a set of related concepts and terms. LSA is known to combat the effects of synonymy (a state in which a word is a synonym for other words) and polysemy (that a word or phrase may have more than one meaning).

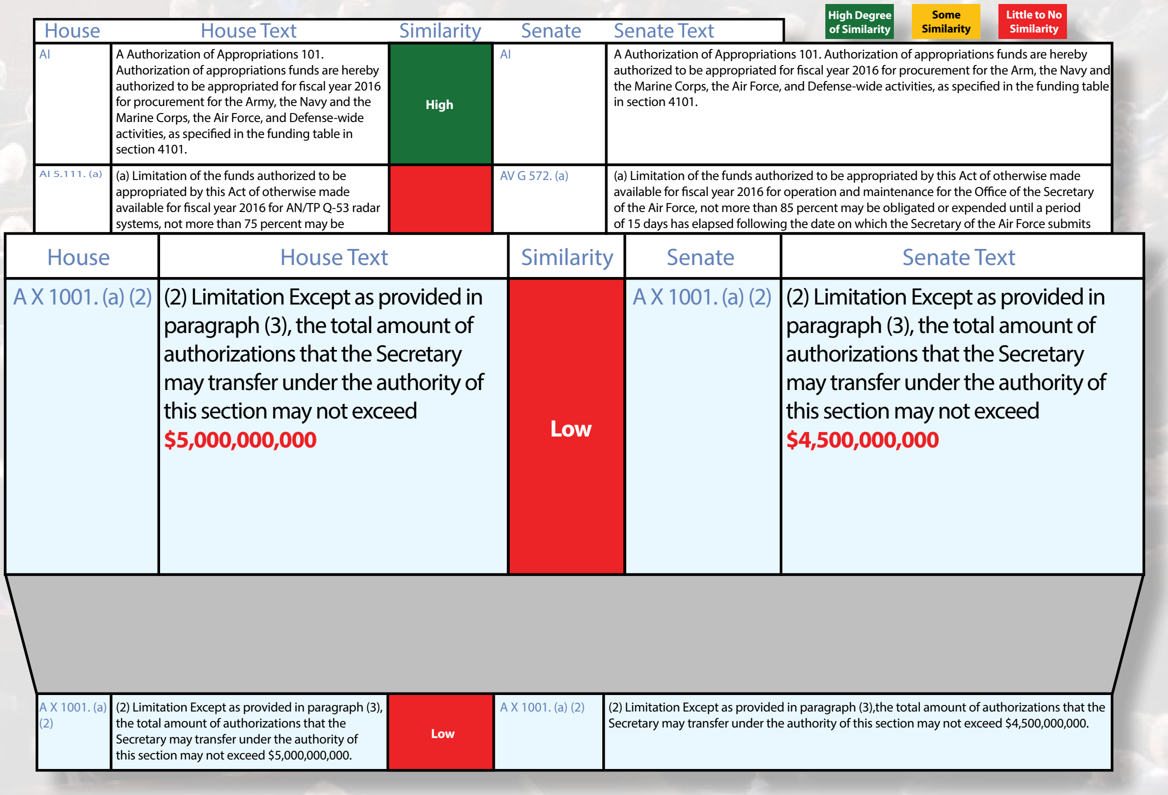
embedded in this space, the team paired off the points, one from the House version and one from the Senate version per pair, starting with the pair that was closest together according to Euclidean distance. At a certain distance, we considered the points too far away from each other to have a relationship. These unpaired points were labeled “no match.”

We then used the point pairings (and the unpaired points) to automatically generate a table. The table was color-coded on a red-yellow-green spectrum, with red indicating a low level of similarity between points, yellow indicating a medium level of similarity, and green indicating a high level of similarity. Figure 1 shows an excerpt of the table, which had 2,982 rows.

Results

The resulting spreadsheets required a human analyst to clean and verify the data. The algorithm sometimes missed connections that it should have made or made unwarranted connections. Overall, however, the algorithm was able, with a high statistical probability, to correctly find sections that were substantially the same. This allowed analysts to concentrate their efforts on the differences between sections.

The algorithm provides substantial time and cost savings for both the analysts and DoD. Because verification is faster than production, analysts require less time to verify or correct an algorithmically produced alignment than to find the same alignment manually. The algorithm



🟩 High Degree of Similarity | 🟨 Some Similarity | 🟥 Little to No Similarity

House	House Text	Similarity	Senate	Senate Text
AI	A Authorization of Appropriations 101. Authorization of appropriations funds are hereby authorized to be appropriated for fiscal year 2016 for procurement for the Army, the Navy and the Marine Corps, the Air Force, and Defense-wide activities, as specified in the funding table in section 4101.	High	AI	A Authorization of Appropriations 101. Authorization of appropriations funds are hereby authorized to be appropriated for fiscal year 2016 for procurement for the Arm, the Navy and the Marine Corps, the Air Force, and Defense-wide activities, as specified in the funding table in section 4101.
AI 5.111. (a)	(a) Limitation of the funds authorized to be appropriated by this Act of otherwise made available for fiscal year 2016 for AN/TP Q-53 radar systems, not more than 75 percent may be	🟥	AV G 572. (a)	(a) Limitation of the funds authorized to be appropriated by this Act of otherwise made available for fiscal year 2016 for operation and maintenance for the Office of the Secretary of the Air Force, not more than 85 percent may be obligated or expended until a period of 15 days has elapsed following the date on which the Secretary of the Air Force submits
A X 1001. (a) (2)	(2) Limitation Except as provided in paragraph (3), the total amount of authorizations that the Secretary may transfer under the authority of this section may not exceed $5,000,000,000.	Low	A X 1001. (a) (2)	(2) Limitation Except as provided in paragraph (3),the total amount of authorizations that the Secretary may transfer under the authority of this section may not exceed $4,500,000,000.

House	House Text	Similarity	Senate	Senate Text
A X 1001. (a) (2)	(2) Limitation Except as provided in paragraph (3), the total amount of authorizations that the Secretary may transfer under the authority of this section may not exceed **$5,000,000,000**	Low	A X 1001. (a) (2)	(2) Limitation Except as provided in paragraph (3), the total amount of authorizations that the Secretary may transfer under the authority of this section may not exceed **$4,500,000,000**

Figure 1. House vs. Senate Language, NDAA 2016

reduces the analysts' role in the initial search for differences – a shift from search and filter to verification and correction.

The time saved can enable the Office of Legislative Affairs to increase throughput without hiring new employees and will allow current employees to focus on tasks that require critical thinking.

References

Black, P.E. 2004. "Euclidean Distance." In *Dictionary of Algorithms and Data Structures,* edited by V. Pieterse and P.E. Black. December 17, 2004. Available from https://xlinux.nist.gov/dads/HTML/euclidndstnc.html. Accessed October 2, 2017.

tfidf.com. http://www.tfidf.com/. Accessed September 26, 2017.

Discovering, Analyzing, and Understanding Improvised Explosive Device Documents

Forrest R. Frank

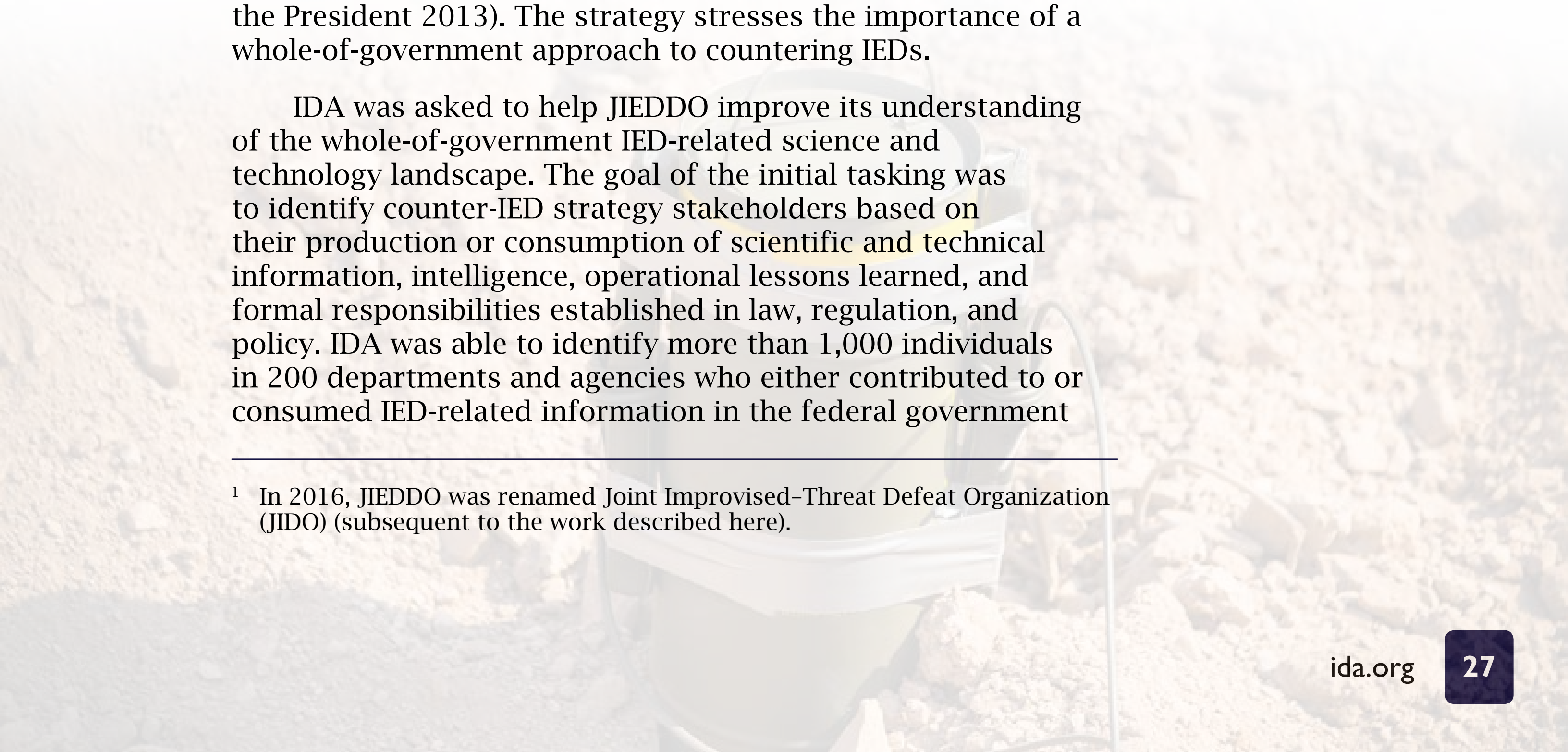
The Challenge: The Joint Improvised Explosive Device Defeat Organization needed a way to improve its understanding of the whole-of-government IED-related science and technology landscape.

Background

The Joint Improvised Explosive Device Defeat Organization (JIEDDO)[1] was established in February 2006 as a joint activity to "focus (lead, advocate, coordinate) all Department of Defense action in support of the Combatant Commanders' and their respective Joint Task Forces' efforts to defeat improvised explosive devices (IED) as weapons of strategic influence" (Department of Defense 2006). To better understand and ultimately defeat adversary IED use against U.S. and coalition forces, JIEDDO sought information from scientific and technical activities overseen by the Joint IED Test Board, individual DoD Components, and other U.S. Government departments and agencies.

To better understand and ultimately defeat adversary IED use against U.S. and coalition forces, JIEDDO sought information from scientific and technical activities overseen by the Joint IED Test Board.

In February 2013, the President released his IED strategy, *Countering Improvised Explosive Devices*, which aimed to expand the Administration's counter-IED focus by building on existing policy and strategy that establish and implement measures to discover, prevent, protect against, respond to, recover from, and mitigate IED attacks and their consequences (Executive Office of the President 2013). The strategy stresses the importance of a whole-of-government approach to countering IEDs.



IDA was asked to help JIEDDO improve its understanding of the whole-of-government IED-related science and technology landscape. The goal of the initial tasking was to identify counter-IED strategy stakeholders based on their production or consumption of scientific and technical information, intelligence, operational lessons learned, and formal responsibilities established in law, regulation, and policy. IDA was able to identify more than 1,000 individuals in 200 departments and agencies who either contributed to or consumed IED-related information in the federal government

[1] In 2016, JIEDDO was renamed Joint Improvised–Threat Defeat Organization (JIDO) (subsequent to the work described here).

alone. However, aligning JIEDDO-produced or JIEDDO-sponsored information with these individuals or organizations proved more difficult.

Applying IDATA to IDA's Research

JIEDDO's scientific and technical information processes supporting IED technical solution test and evaluation had assumed a wartime mentality. JIEDDO's focus was on delivering counter-IED capability (i.e., materiel solutions and some related tactics, techniques, and procedures (TTP)) to warfighters. Delivering documentation of scientific, engineering, and testing activities to JIEDDO or the Defense Technical Information Center (DTIC) was deferred in favor of fielding counter-IED capabilities from 2006 until 2013. As a result, JIEDDO lacked understanding of its trove of scientific and technical data and reports that could support the whole-of-government approach to countering IEDs.

In late 2013, IDA was asked to capture scientific and technical data and reports distributed to JIEDDO's information systems, as well as accessible data produced by JIEDDO-funded test and evaluation activities. IDA was also asked to develop a process prototype that would make this information and data available for indexing, search, and retrieval. We thought that the IDATA capability, even in its preliminary design, would be a potential partial solution.

The first challenge was locating and retrieving data. We collected more than 7,000 documents from multiple classified and unclassified JIEDDO information systems. We also collected more than 1,000 documents from information systems operated by the Army and the Navy, where JIEDDO had provided funds, tasking, or data for use in tests, evaluations, and experiments that resulted in the documents stored on those systems.

The next challenge was to identify and characterize the data. We determined that the data would be held on a standalone SECRET//NOFORN system due to its sensitivity when aggregated. Running in a system-high mode on a standalone, classified, power-gaming-type computer allowed us to exercise IDATA in a new environment. The initial version of IDATA was able to characterize documents by source, classification level, and handling requirements. It also automatically generated keywords based on an algorithmic assessment of each document's content. As such, we were able to quickly triage document content and make decisions regarding further processing. In some cases, we ran IDATA against a document multiple times to extract additional keywords or obtain keyword counts to measure importance based on the number of individual keywords, keyword frequency, and trends regarding keyword location within the document.

A second iteration of the IDATA tool set was prepared for this task. This improved capability was equipped to aggregate and count keywords to help generate statistics illustrating the use of certain keywords within a document or set of documents. We could, for example, run IDATA against a set of documents from an information system (e.g., a Navy-housed database) to find metadata and other content providing insight into test and evaluation processes, successes, and challenges that might have otherwise become apparent only by reading several hundred individual documents.

This ability to look at the frequency and importance of keywords is roughly analogous to the early use of machine translation of Russian and Chinese scientific reports. The data are indicative of content and help subject matter experts select specific documents or data sets for further analysis.

The improved capability also delved deeper into the raw information and extracted text information embedded in analog data sets. For example, IDATA helped identify important embedded scientific and technical data by deciphering text descriptions in graphics; the titles of figures recorded as JPGs, PNGs, or TIFF images; and descriptions accompanying analog audio recordings.

Task Results

IDA researchers demonstrated that IDATA can quickly process and index thousands of documents and discover important keywords without extensive human intervention and laborious document review. IDATA provides a variety of options for collecting a wide array of information and making it available to DoD and other counter-IED organizations if JIDO (Joint Improvised-Threat Defeat Organization, the successor organization to JIEDDO) chooses to build its own information storage, retrieval, and dissemination capability.

Alternatively, if JIDO chooses to rely on DTIC as the repository of all JIEDDO information, including digital and analog test and evaluation data, IDATA's demonstrated ability to automatically generate keywords could be used to facilitate the completion of DTIC's documentation requirements.

References

Department of Defense. 2006. "Joint Improvised Explosive Device Defeat Organization (JIEDDO)." *DoD Directive 2000.19E*. Washington, DC: Department of Defense, February 14, 2006.

Executive Office of the President. February 26, 2013. "Countering Improvised Explosive Devices." Washington, DC: Executive Office of the President.

Use of IDATA Capabilities for Social Media Analytics

Thi Uyen Tran and Daniel Nakada

The Challenge: The Defense Threat Reduction Agency needed to understand how well a biosurveillance application finds relevant disease information on social media.

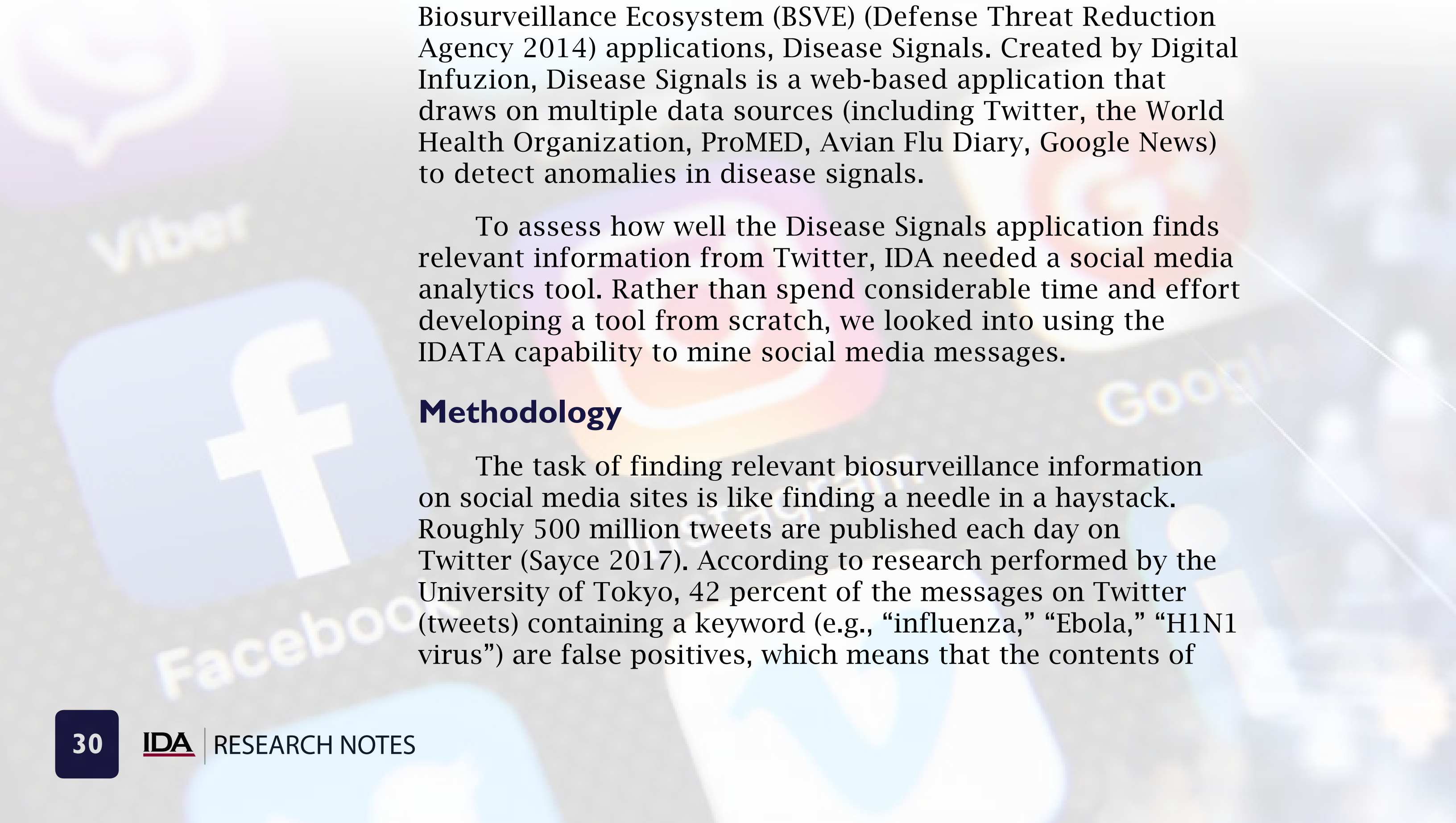
Social media sites produce information that research communities can use to improve responses to national security and public health problems, such as measuring public anxiety after a natural disaster.

Background

In recent years, the emergence of social media sites such as Facebook, Twitter, Snapchat, and LinkedIn has fundamentally shifted the way people communicate and share information. Today, political views, religious beliefs, and even personal health status can be transmitted easily at near-real-time speed. This phenomenon produces a wealth of information that research communities can use to improve responses to national security and public health problems, such as measuring public anxiety after a natural disaster (Doan, Ho Vo, and Collier 2011), detecting an earthquake through the "social sensor" (Sakaki, Okazaki, and Matsuo 2010), monitoring bribery or violence during an election (Draxler 2014), or detecting and tracking infectious disease outbreaks.

IDA was asked to assist the Defense Threat Reduction Agency (DTRA) with its evaluation of one of the Biosurveillance Ecosystem (BSVE) (Defense Threat Reduction Agency 2014) applications, Disease Signals. Created by Digital Infuzion, Disease Signals is a web-based application that draws on multiple data sources (including Twitter, the World Health Organization, ProMED, Avian Flu Diary, Google News) to detect anomalies in disease signals.

To assess how well the Disease Signals application finds relevant information from Twitter, IDA needed a social media analytics tool. Rather than spend considerable time and effort developing a tool from scratch, we looked into using the IDATA capability to mine social media messages.



Methodology

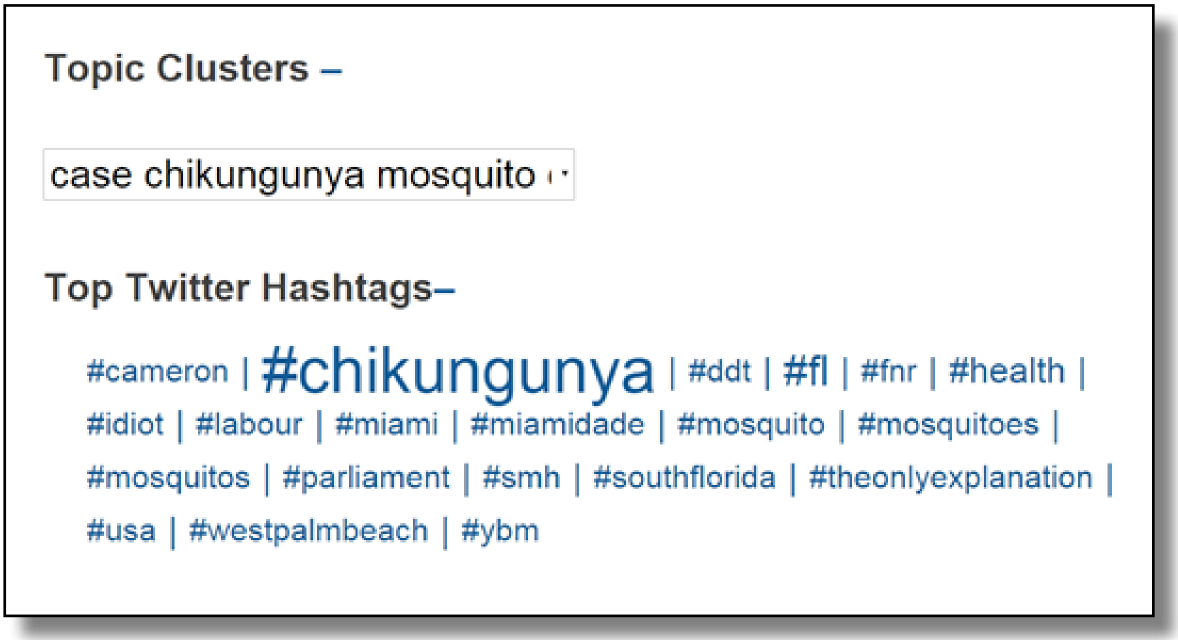
The task of finding relevant biosurveillance information on social media sites is like finding a needle in a haystack. Roughly 500 million tweets are published each day on Twitter (Sayce 2017). According to research performed by the University of Tokyo, 42 percent of the messages on Twitter (tweets) containing a keyword (e.g., "influenza," "Ebola," "H1N1 virus") are false positives, which means that the contents of these tweets are irrelevant to the topics of interest (i.e., influenza, Ebola, or H1N1) (Aramaki, Maskawa, and Morita 2011). The objective of mining social media messages is to reduce the social noise as much as possible to minimize false positives, which can lead to false alarms of an emerging disease.

The other challenge was to discover new topics that arise without prior knowledge, e.g., a new virus breakout or a natural disaster. We employed IDATA's capabilities to address this problem. Although IDATA was not originally intended for social media analysis, it is designed to be highly customizable and extensible. In this case, IDATA was easily extended to ingest tweets, extract hashtags from those tweets, and display the resultant trends.

In 2014, IDA began to feed the first set of tweets into IDATA. To narrow down the scope of the topics, we limited the search to messages that contained a set of keywords related to health, such as "fever," "flu," "influenza," "virus," "infection," "measles," "H1N1," and "pneumonia." IDATA key phrase extraction quickly revealed a high degree of false positives generated by health-related keywords such as "World Cup" (e.g., "World Cup fever") and "Brazil." This illustrates the challenge: A search for "fever" led to posts about the World Cup. We had to determine the best way to use IDATA features to adjust to humans' semantic ambiguity – that is, humans' tendency to ascribe meaning and purpose to words that may differ from the words' original meaning and context (such as *fever* meaning *high temperature* versus *World Cup fever*).

Results

IDATA's underlying analytics, such as topic discovery and entity extraction, in addition to the interactive interface, helped sift through the noise to zero in on buried signals. For example, as shown in Figure 1, we discovered an emerging health topic related to the mosquito-



Topic Clusters –

case chikungunya mosquito

Top Twitter Hashtags–

#cameron | #chikungunya | #ddt | #fl | #fnr | #health | #idiot | #labour | #miami | #miamidade | #mosquito | #mosquitoes | #mosquitos | #parliament | #smh | #southflorida | #theonlyexplanation | #usa | #westpalmbeach | #ybm

Figure 1. IDATA Discovers a New Topic: Chikungunya

borne Chikungunya virus because an individual in Florida was reported to be infected at the time.

IDATA was not originally designed for social media analysis, and its algorithms (e.g., topic models) are not specifically optimized for microblog data, which contain emoticons and a shorthand, colloquial language style. Despite this, IDATA provided good results on this task when applied to Twitter data. Moreover, IDATA's high degree of extensibility made it easy to customize for social media.

References

Aramaki, E., Maskawa, S., and Morita, M. (2011). "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter." *Proceedings from the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, July 27-31, 2011.

Defense Threat Reduction Agency. (2014). "The Biosurveillance Ecosystem (BSVE)." http://www.dtra.mil/Portals/61/Documents/bsve-fact-sheet_draft_05-01-2014_pa-cleared-distro-statement.pdf.

Doan, S., Ho Vo, B., and Collier, N. (2011). *An Analysis of Twitter Messages in the 2011 Tohoku Earthquake*. Paper presented at the eHealth 2011 conference, Malaga, Spain.

Draxler, B. (2014). "How Tweets Can Save Lives." *Popular Science*. September 18, 2014 https://www.popsci.com/article/how-tweets-can-save-lives.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors." *Proceedings from the 19th International Conference on World Wide Web*, Raleigh, North Carolina, April 26-30, 2010: 851-860.

Sayce, D. (2017). "Number of Tweets per Day?" *David Sayce*. https://www.dsayce.com/social-media/tweets-day/.

Conclusion

Michelle Albert

The IDATA Team

Recently, the open-source community and the commercial marketplace have exploded with data triage and discovery capabilities. Most of these tools, however, are not designed for DoD or federal government use. Commercially available tools rarely provide data pre-processing or post-processing. They have no knowledge of the context in which DoD and other government agencies operate or where the data come from.

The IDATA capability, on the other hand, uses cutting-edge open-source libraries and a modular approach that enables it to be customized to suit a sponsor's needs and work flow. The other major part of IDATA's value is the IDA analysts, subject matter experts, and researchers who work with the capability. They come from a variety of backgrounds – academia, industry, military – and have diverse experience and education. They include the expected computer scientists and mathematicians, but also include the expertise of English majors, political scientists, economists, and physicists. Just as non-correlated algorithms can compensate for each other's weaknesses when used together, a well-rounded team can address myriad disparate tasks and issues.

This approach and the IDATA team's results have been recognized by IDA and the scientific community at large. The team has published multiple papers and has received the IDA Welch Award, IDA's W.Y. Smith Award, and awards from the Association of Enterprise Information (AFEI, an affiliate of the National Defense Industrial Association) for Automated Information Triage for Rapid Decision Making, in 2015, and State Cyber Operations Framework, in 2016.

Future Opportunities for the IDATA Capability

The IDATA team continues to grow, evolve, and incorporate new tools and techniques to stay on the cutting edge of natural language processing and

2016 Larry D. Welch Award for Best External Publication winners: Andrew Wan, Arun Maiya, and Dale Visser (left to right).

Miranda Seitz-McLeese accepts the W.Y. Smith Award.

The IDATA Team receives the AFEI Award.

From left to right: Andrew Wan, Edna Jordan, Anna Vasilyeva, Thomas Barth, Andrew Ferguson, Tristian Weir, Dave Chesebrough (AFEI), Laura Odell, Miranda Seitz-McLeese, Cameron DePuy, and Corbin Fauntleroy provide sponsors with tailored, non-biased, and actionable results.

The most fertile ground for the IDATA capability lies in areas with large quantities of public data, especially the growing amount of information available online that is released both by private individuals on social media and by the U.S. Government as part of the Open Data Policy.[1]

Sentiment analysis could provide a new way for sponsors to look at this available data. Tracking connections via LinkedIn, geotags on Facebook, or hashtag use on Twitter may provide sponsors with valuable insight. Recruiters, for example, could use these data to more easily find and target potential recruits with desired skill sets.

Also, the Office of Legislative Affairs could track Twitter posts from a particular location to help anticipate the concerns of a particular Congress member's constituents. There are algorithms that would work with efforts of this kind, although they have not been fully employed in a national security context. The IDATA capability's modular nature makes it fully compatible with them.

Another avenue for potential growth comes from the discovery that existing entity-extraction algorithms, which attempt to extract the names of people and organizations from natural language documents, do not adapt well to the national security context. Training and modifying these algorithms would make the task of sifting through and sorting a large amount of government documents less daunting. For example, users could filter for documents containing a particular office's name or filter for memoranda signed by a particular official. Contracts could also be filtered by company, which would potentially enable DoD to consolidate purchases and eliminate duplication.

These are just a few examples. Many more may become available as research continues in this area and analysts apply new techniques.

Summary

By covering a wide range of topic areas, IDA has proven that the capability underlying the text analytics concept can be rapidly customized to produce results in a relatively short time. Once sponsors have the opportunity to use automated information triage to solve a problem, they quickly see the benefit and bring in other problems to solve. This approach has also given sponsors in DoD and other federal agencies the opportunity to reduce costs due to the IDATA capability's ability to minimize the time needed to find, search through, and analyze information across a variety of documents and file types. It has also enabled DoD and other federal agencies to find and process existing information, eliminating duplicative activities that result from organizations' inability to find or manipulate data from earlier efforts.

[1] Executive Order (EO) 13642, "Making Open and Machine Readable the New Default for Government Information," states that making information accessible and usable can promote job growth, innovation, and scientific discovery. It establishes a default in which data are released to the public whenever possible and legally permissible.

Reference

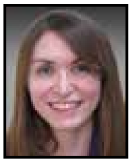
The White House. Executive Order 13642, *Making Open and Machine Readable the New Default for Government Information.* May 9, 2013.

Postscript

The articles in this *IDA Research Notes* describe examples of how the IDATA capability has been used to solve real-world problems. Those who wish to know more about the research behind IDATA should review the following papers:

- *Mining Measured Information from Text* (published at SIGIR '15) (https://doi.org/10.1145/2766462.2767789)
- *A Framework for Comparing Groups of Documents* (published at EMNLP '15) (https://www.ida.org/idamedia/Corporate/Files/Publications/IDA_Documents/ITSD/2015/D-5543.pdf)
- *Topic Similarity Networks: Visual Analytics for Large Document Sets* (published at IEEE BigData '14) (https://ieeexplore.ieee.org/document/7004253/)
- *Exploratory Analysis of Highly Heterogeneous Document Collections* (published at KDD '13) (https://doi.org/10.1145/2487575.2488195)
- *Supervised Learning in the Wild: Text Classification for Critical Technologies* (published at IEEE MILCOM '12) (https://ieeexplore.ieee.org/document/6415660/)

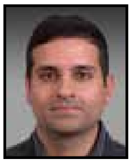
Contributors



Ms. Michelle Albert is a Research Associate in IDA's Information Technology and Systems Division. She holds a Master of Arts in journalism from the University of Missouri.

Ms. Katharine Burton is an Adjunct Research Staff Member in IDA's Information Technology and Systems Division. She holds a Master of Arts in administrative science from George Washington University and a Master of Science in national security strategies from the National Defense University.

Dr. Forrest Frank is a Consultant in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in political science from Stanford University.

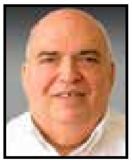


Dr. Arun Maiya is a Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in computer science from the University of Illinois at Chicago.

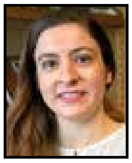
Dr. Daniel Nakada is a former Research Staff Member in IDA's System Evaluation Division. He holds a Doctor of Philosophy in electrical engineering from the Massachusetts Institute of Technology.

Mr. James O'Connor is a former Research Assistant in IDA's Information Technology and Systems Division. He holds a Master of Science in mechanical engineering from Worcester Polytechnic Institute.

Ms. Laura Odell is an Assistant Director in IDA's Information Technology and Systems Division. She holds a Bachelor of Science in electrical engineering from the University of Minnesota, a Master of Science in contract acquisition management, and a Master of Science in management from the Florida Institute of Technology.

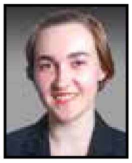


Dr. Robert Rolfe is an Adjunct Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in physics from the University of California, Los Angeles.



Ms. Miranda Seitz-McLeese is a Research Associate in IDA's Information Technology and Systems Division. She holds a Master of Science in applied mathematics and statistics from Georgetown University.

Ms. Thi Uyen Tran is a Research Staff Member in IDA's System Evaluation Division. She holds a Bachelor of Science in electrical engineering from George Mason University.



Ms. Anna Vasilyeva is a former Research Associate in IDA's Information Technology and Systems Division. She holds a Master of Science in aeronautics and astronautics from the Massachusetts Institute of Technology and a Master of Science in technology policy from the Judge Business School, University of Cambridge.

Dr. Dale Visser is a Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in physics from Yale University.

Dr. Andrew Wan is an Adjunct Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in computer science from Columbia University.

Past Issues

Challenges in Cyberspace

- Cyberspace – The Fifth and Dominant Operational Domain
- Transitioning to Secure Web-Based Standards
- Information Assurance Assessments for Fielded Systems During Combat Command Exercises
- Supplier-Supply Chain Risk Management
- Internet-Derived Targeting: Trends and Technology Forecasting
- Training the DoD Cybersecurity Workforce

Multidisciplinary Research for Securing the Homeland – IDA and DHS: Beyond 15

- Countering Terrorism One Technology at a Time
- Does Imposing Consequences Deter Attempted Illegal Entry into the United States?
- Improving Shared Understanding of National Security and Emergency Preparedness Communications
- Foreign Counter-Unmanned Aerial Systems: Developments in the International Arms Market
- Operationalizing Cyber Security Risk Assessments for the Dams Sector
- Understanding the Juvenile Migrant Surge from Central America
- Implementing a Roadmap for Critical Infrastructure Security and Resilience
- Baselining: Application of Qualitative Methodology for Quantitative Assessment
- Analysis, Analysis Practices, and Implications for Modeling and Simulation
- Test and Evaluation for Reliability

Acquisition, Part 1: Starting Viable Programs

- Defining Acquisition Trade Space Through "DERIVE"
- Supporting Acquisition Decisions in Air Mobility
- Assessing Reliability with Limited Flight Testing
- Promise and Limitations of Software Defined Radios
- Implications of Contractor Working Capital on Contract Pricing and Financing
- The Mechanisms and Value of Competition
- Early Management of Acquisition Programs

Acquisition, Part 2: Executing and Managing Programs

- Cost Growth, Acquisition Policy, and Budget Climate
- Improving Predictive Value of Poor Performance

Past Issues Continued

- Root Cause Analysis of VTUAV Fire Scout's Nunn-McCurdy Breach
- Evaluating Solid Rocket Motor Industrial Base Consolidation Scenarios
- Managing Supply Chain Cyber Risks To DoD Systems and Networks
- Looking Back at PortOpt: An Acquisition Portfolio Optimization Tool
- Predicting the Effect of Schedule on Cost
- Recent Developments in the Joint Strike Fighter Durability Testing

Test and Evaluation: Statistical Methods for Better System Assessments

- Assessing Submarine Sonar Performance Using Statistically Designed Tests
- Applying Advanced Statistical Analysis to Helicopter Missile Targeting Systems
- Tackling Complex Problems: IDA's Analyses of the AN/TPQ-53 Counterfire Radar
- Improving Reliability Estimates with Bayesian Hierarchical Models
- Managing Risks: Statistically Principled Approaches to Combat Helmet Testing
- Validating the Probability of Raid Annihilation Test Bed Using a Statistical Approach

Technological Innovation for National Security

- Acquisition in a Global Technology Environment
- Lessons on Defense R&D Management
- Commercial Industry R&D Best Practices
- Strengthening Department of Defense Laboratories
- Policies of Federal Security Laboratories
- The Civilian Science and Engineering Workforce in Defense Laboratories
- Technology Transfer: DoD Practices

Security in Africa

- Trends in Africa Provide Reasons for Optimism
- China's Soft Power Strategy in Africa
- Sudan on a Precipice
- A New Threat: Radicalized Somali-American Youth
- Chinese Arms Sales to Africa

IDA RESEARCH NOTES

© Institute for Defense Analyses

4850 Mark Center Drive • Alexandria, VA 22311-1882
ida.org

@IDA_org