MITRE | Center for Data-Driven Policy

SERIES Number 10

INTELLIGENCE AFTER NEXT

WHAT ARE WE WAITING FOR? TODAY'S TECHNOLOGIES CAN IMPROVE THE BUSINESS OF ALL-SOURCE INTELLIGENCE ANALYSIS NOW

by Michael Maskaleris

Approved for Public Release; Distribution Unlimited. Public Release Case Number 21-3035

Today's technologies can help all-source analysts now

The Intelligence Community (IC) is focused on developing and deploying future cutting edge technical capabilities—including artificial intelligence, machine learning, and deep learning—to support all-source intelligence analysis. In the interim, existing technologies, especially human language technology and natural language processing capabilities, available in the commercial world for years, should be deployed to ease analysts' workload while making them more efficient until these future systems are operational.

Improvements in the following areas would have the greatest impact on the day-to-day work of an analyst:

- **Information retrieval.** Deployment of non-Boolean search capabilities and a desktop search application will provide analysts with the means to use other search strategies to find needed information, as well as free analysts from the frustration of spending considerable time searching for data saved in electronic folders and drives.
- **Information filtering.** Integration of information extraction and summarization with existing agency information retrieval tools will assist analysts with knowledge discovery by reviewing large numbers of search returns as well as identifying unexpected relationships.

Ultimately analysts need a capability that will decrease the time it takes to search for relevant information and provide them with more time to think, analyze, and write. Such a capability should not only provide federated search of IC data repositories, but also automatically extract and filter information, as well as push data in the correct formats so that they can be visualized in other applications.

Helping analysts think, analyze, and write

Over the years, national-level intelligence agencies have attempted many efforts to improve or modernize all-source intelligence analysis using technology. However, few of these programs reach fruition and are currently deployed on classified networks. Where there have been successes, they have been in support of relatively small groups of all-source intelligence analysts assessing focused issues. Their solutions are not scaled to or adopted by the greater analytic enterprise. As a result, analysts continue to lack capabilities for knowledge discovery that have been commercially available for years. Most striking is a failure to deploy human language technology (HLT) and natural language processing (NLP) capabilities to assist analysts in information retrieval and filtering. Using these existing technologies could greatly ease analysts' work while they wait for the operationalization of artificial intelligence, machine learning, and other future capabilities. However, deploying better information retrieval and filtering capabilities may require replacing some current information retrieval tools linked to agency data repositories because these tools have not kept pace with those available in the commercial world.

Information retrieval

The current environment

All-source intelligence analysts at the national level acquire classified information by conducting manual queries of data repositories, generally containing unstructured text reports, maintained by each separate intelligence agency. Searching often requires mastering a different information retrieval tool for each repository, some of which are primitive. Analysts also can access and research open-source intelligence (OSINT) information on the Internet via the Non-classified Internet Protocol Router Network (NIPRNET) using unclassified workstations. However, transferring OSINT one file at a time from the NIPRNET is a time-consuming process and somewhat limited by file-type restrictions that can

RATHER THAN WAIT FOR THE NEXT GENERATION OF ANALYTIC TECHNOLOGIES AND TOOLS, ANALYSTS NEED IMPROVED CAPABILITIES NOW TO RETRIEVE, FILTER, AND DISCOVER CRITICAL DATA THAT WILL DECREASE THE TIME IT TAKES TO FIND NEEDED INFORMATION

prevent moving media and zipped files. Videos, imagery, photographs, stock maps, and other non-text data are stored in separate repositories, and all require mastery of a repository-specific search tool.

Analysts have the capability to save data they wish to retain in many of the information retrieval tools they have used to find the information. However, there is no easy way to aggregate saved data between repositories. Consequently, analysts usually save this information in personal electronic or shared electronic folders that are not easily searchable due to security limitations that may index only the file name and type and not the full text contained within the file.

Today's technologies and analytic solutions

Desktop search

Of highest priority is the deployment of a desktop search capability. These tools search within a user's own computer files as opposed to searching the Internet or Intelink. They are designed to find information in saved electronic files, including web browser history, email archives, text documents, sound files, images, and video. To ensure that their most important information is readily available when needed, almost all analysts print large numbers of documents so that they can store them "hard copy." Deployment of an optimized desktop search application, such as Copernic Desktop Search, would significantly solve this problem and reduce analysts' workload and printing costs.

Non-Boolean search

Boolean searches use individual words and phrases—"keywords"—to find information that matches these terms. Boolean queries inherently limit knowledge discovery, since analysts build their searches using terms they already know. In addition, creating Boolean queries to find needed information without having to sort through many false positive returns is often an iterative and time-consuming process.

Non-Boolean search using NLP works differently from keyword searches, as it tries to capture the meaning of a user query and answer the question instead of merely matching the keywords. Deploying non-Boolean search capabilities integrated with the IC's data repositories could improve analysts' knowledge discovery and greatly reduce the time expended to find needed information.[1] Some existing non-Boolean search technologies that could be deployed include:

- **Question Answering (QA)** uses complex NLP techniques to retrieve answers to a wide range of questions posed in natural language. Analysts are already familiar with using QA to find information while working at home. This technology has been available in many Internet search capabilities for years, such as Google, Bing, and DuckDuckGo. Using QA would greatly assist analysts since their searches are often based on answering an underlying question. In addition, posing a question can be accomplished much more quickly than laboriously building a complex Boolean query.
- **Concept Search** is an automated information retrieval method that searches electronically stored unstructured text. It performs queries using specific terms that are then expanded to other terms that express a similar idea. Searching with concepts rather than keywords improves precision by eliminating unintended senses of the search terms. It also facilitates data discovery by potentially displaying unexpected relationships between the different concepts. In addition, concept search can provide powerful extensions based on relationships between concepts.[2 3 4]

- **Reverse Image Search** allows a user to upload a graphic to find visually similar images and obtain relative information about it, including its name, objects or places in it, and other associated metadata instead of entering a text-based keyword.[5] Such a capability would greatly reduce the time necessary for finding needed graphics instead of attempting to do so using text-only searches.

Information filtering

The current environment

Current information retrieval applications temporally display query results as rows of "hits" containing information such as Product ID, Producing Agency, Creation Date, and Report Title/Message Subject Line. Some information retrieval tools provide limited filtering of query results but only of existing tags that are associated with the individual report or message such as Producing Agency, Classification, Topic Country, and Intelligence Function Code (IFC). None of the tools use NLP to filter the full text of a document. Consequently, analysts do not have a way to prioritize their time by reviewing the most relevant query returns first. Instead, they must scan through rows of titles and use their experience to guess which documents are most likely to be of interest.

When analysts find a result that they wish to keep, current information retrieval tools may provide the capabilities to save it within the application, export it as a PDF file to be saved on an electronic drive, or print it. Analysts often print their most important information due to the difficulties in finding files stored in electronic folders already discussed.

Technical capabilities available for deployment

NLP has several technologies that, if integrated with existing IC information retrieval tools, would significantly assist in prioritizing query results and assisting knowledge discovery. They include:

- **Document Clustering:** detection of topics within a data source and grouping together documents on similar topics. Integrating document clustering capabilities to existing IC information retrieval tools would greatly assist analysts in prioritizing the review of large volumes of information within the context of search results, as well as facilitate knowledge discovery. With concepts displayed either visually or as lists, analysts would be able to quickly find documents of most interest to them. They would also be able to discover reports of potential interest based on the "closeness" of documents within a cluster and between clusters, as well as potential unexpected relationships between the clusters.
- **Summarization:** the process of distilling the most important information from sources and producing an abridged version as either an abstract or an extract. Adding a summary of the contents of each "hit" on the list of query returns would allow analysts to quickly scan their results and identify reports of most interest.
- **Information Extraction:** the identification of specific semantic elements expressed within a linguistic utterance, including text, speech, and video. Historically the most frequently examined types of "semantic elements" extracted include entities (e.g., people, organizations, locations), their various referring expressions (names, descriptions or nominals, and pronouns), relationships among entities and other values, and events. There are as many different types of analyses possible as there are categories of meaning that might be extracted, and research continues to explore new types of information that might yield better automated processes.[6]
- **Entity Extraction:** also known as entity name extraction, or named entity recognition, an information extraction technique that refers to the process of identifying and classifying key elements from text into pre-defined

categories. The technique helps transform unstructured data to data that is structured, and therefore machine readable and available for standard processing that can be applied for retrieving information, extracting facts, and question answering.[7]

- **Location Extraction:** identifies geospatial information within a data source. Most location extraction relies on recognizing geocoordinates. However, gazetteers and other resources can be used to discover place names and then apply coordinates to these identified locations. In some cases, disambiguation is used by placing these place names in the context of the location (i.e., Paris, France versus Paris, Texas).[8]
- **Event Extraction:** identifies actions or activities within a data source, most commonly based on a pre-defined ontology of events that can then be recognized. NLP can also recognize verbs within sentence construction to extract activity.[9]
- **Relationship Extraction:** identifies named relationships between entities in text. Semantic role labeling is a closely related task. Such extraction usually uses a list of pre-defined relationships, such as "son of," "works with," etc. to determine relationships between two entities. These tools can also use entity co-occurrence with NLP to extract relationships, such as subject-link-object constructions in a sentence.[10]

Integrating information extraction capabilities to existing IC information retrieval tools would greatly assist analysts in prioritizing the review of large volumes of information, and in knowledge discovery. Analyst queries would result in organized lists of people, organizations, locations, events, relationships, and other data that has been extracted from the search results, allowing them to quickly access information concerning entities of most interest to them. NLP-enhanced lists of entities could also assist knowledge discovery by identifying unexpected links between entities. Some of these entities could be exported and displayed in other applications, such as geospatial, link, and temporal analysis tools, providing different views of the data and facilitating knowledge discovery.

Let's stop waiting

Employing the capabilities described in this paper for IC analysts will provide federated search of IC data repositories, automatically extract and filter information, and push data in the correct formats so that they can be visualized in other applications. Rather than wait for the next generation of analytic technologies and tools, analysts need improved capabilities now to retrieve, filter, and discover critical data that will decrease the time it takes to find needed information, thus providing them with additional time to do their critically important missions: think, analyze, and write.

Notes

1 O'Reilly (n.d.), "Natural Language Processing," retrieved on January 12, 2021.
Available: https://www.oreilly.com/library/view/introduction-to-search/9780596809546/ch01.html

2 V. Prajapati (n.d.), "What is the difference between a conceptual search and semantic search?," *Quora*, retrieved on January 12, 2021.
Available: https://www.quora.com/What-is-the-difference-between-a-conceptual-search-and-semantic-search

3 M. Maybury, "State-of-the-art tools for more efficient information discovery and analysis."
Presented at the Society for Competitive Intelligence Professionals Annual Conference in Orlando, Fla., April 26-29, 2006. Approved for MITRE public release. Retrieved on January 12, 2021.
Available: https://communityshare.mitre.org/sites/CHIP/MITRE%20Papers%20and%20Presentations/State-of-the-Art%20Tools%20for%20More%20Efficient%20Information%20Discovery%20and%20Analysis.pdf#search=%22State%2Dof%2Dthe%2Dart%20tools%20for%20more%20efficient%20information%20discovery%20and%20analysis%22

4 Relativity.com (n.d.), "Concept Searching," retrieved on September 21, 2021.
Available: https://help.relativity.com/RelativityOne/Content/Relativity/Analytics/Concept_searching.htm

5 SmallSetoTools (n.d.), "What Is Reverse Photo Search and How Image Search Works," retrieved on January 12, 2021.
Available: https://smallseotools.com/reverse-image-search/

6 The MITRE Corp (n.d.), "Information Extraction," ATS Lab," retrieved on January 14, 2021.
Available: https://mitrepedia.mitre.org/index.php/Information_Extraction

7 expert.ai, January 16, 2020, "Entity Extraction: How Does It Work?," retrieved on January 14, 2021.
Available: https://www.expert.ai/blog/entity-extraction-work/#:~:text=Entity%20extraction%2C%20also%20known%20as,text%20into%20pre%2Ddefined%20categories

8 The MITRE Corp. (n.d.), "Location Extraction," ATS Lab, retrieved on January 14, 2021.
Available: https://mitrepedia.mitre.org/index.php/Information_Extraction

9 The MITRE Corp. (n.d.), "Event Extraction," ATS Lab, retrieved on January 14, 2021.
Available: https://mitrepedia.mitre.org/index.php/Information_Extraction

10 The MITRE Corp. (n.d.), "Relationship Extraction," ATS Lab, retrieved on January 14, 2021.
Available: https://mitrepedia.mitre.org/index.php/Information_Extraction

Appendix A

Suggested prioritized functional requirements for information retrieval and filtering

This appendix suggests a list of functional requirements to support all-source intelligence analysis, including some capabilities that are not discussed in this paper. Of highest priority is the deployment of a desktop search capability that will free analysts from the frustration of spending an inordinate amount of time searching for saved data. The integration of filtering capabilities—information extraction and summarization—with existing agency information retrieval tools to help analysts prioritize their review of large numbers of search returns and facilitate knowledge discovery is also critical. Preferably, these tools would be integrated with the IC's federated search capability. The deployment of these capabilities could be accomplished in the short-to-medium term.

Capability	Priority
Desktop Search–Index, normalize, and query structured and unstructured data saved in multiple electronic folders and drives including databases, email, HTML, XML, Microsoft Office product formats, PDF, etc.	High
Federated Search–Index, normalize, and perform complex Boolean queries across multiple, massive, geographically separated structured and unstructured data sources, including databases, email, electronic message traffic, HTML, XML, Microsoft Office product formats, PDF, etc. Results from each source should be merged to provide a single integrated list of results, with duplicate "hits" eliminated, and the results sortable by relevancy, source, classification, and data time group/time stamp.	High
Features	
Date Ranges–Perform complex Boolean queries within user defined data ranges.	High
Saved and Shareable Searches–Save queries and allow them to be shareable with other users.	High
Key Word Highlighting–When viewing a query "hit," highlight key words or concepts that resulted in the query being returned.	High
Automatic Query Expansion–Perform query expansion to include word stemming, terms that are phonetically or typographically similar, SOUNDEX, misspellings, synonyms, and wildcards, as well as use pre-and/or user-defined alias lists.	High
Query Result Prioritization–From a list of query results, the user should be able to arrange "hits" based on percentage of likelihood that the "hit" meets query parameters; from most recent to least recent (based on DTG or when the "hit" was added to the data source); unread (not opened and viewed) "hits"; and new "hits" from last time the query was run.	High
Information Extraction/Categorization–From a list of query results, extract entities and categorize them as, at a minimum, people, organizations, countries/locations, equipment/artifacts, phone numbers, email addresses, dates, addresses, locations, and geographic coordinates. Display these extracted terms as part of the query results list and allow a user to "drill down" by selecting one of these extracted entities.	High

Appendix A (cont'd.)

Features	Priority
Document Summarization–Summarize each query "hit" and display these summaries as part of the query return list.	High
Alerts–Monitor data sources based on user-defined queries/"profiles" of interest and provide alerts via Outlook email when data on these terms are added or changed.	High
Scheduled Queries–Query connected data sources on a user-defined, scheduled basis.	High
Concept Search and Extraction–Classify and display query results based on automatically generated and user-created categorizations, ontologies, and/or taxonomies. Display concept or topic "maps" to include associated relationships based on these categorization and taxonomic results.	Medium
Document Clustering–Process query results to determine the distinguishing words or "topics" within each document or data record, based upon statistical measurements of word distribution, frequency, and co-occurrence with other words. Use these distinguishing words to create a mathematical signature or vector for each document in the collection and then do a rough similarity comparison of all the signatures and vectors to create cluster groupings. Compare the clusters against each other for similarity and arrange them in high-dimensional space so that similar clusters are located close together. Flatten the high-dimensional arrangement of clusters down to a comprehensible two or three dimensions—trying to preserve a picture where similar clusters are located close to each other, and dissimilar clusters are located far apart. Link associated documents/data records to their associated cluster.	Medium
Question and Answer Query–Perform queries that are replies to questions (such as What is the importance of person X? What transactions have countries X and Y conducted with each other over the last year?).	Medium
View Most Relevant Information–When a query "hit" is being viewed, it should "open" up to the most relevant portion of the "hit" based on the search terms.	Low

Author

Michael Maskaleris is a principal intelligence analyst at MITRE with 45 years of experience as an intelligence analyst and manager of analysts, and supporting analytic projects. A retired U.S. Army Officer, he has served in intelligence positions from battalion to the national levels, including three tours at DIA. Since his retirement from active duty in 1999, he has worked for MITRE as an all-source intelligence analyst, as a subject matter expert on the Russian armed forces, and developing and training analysts to use new analytic methodologies, data sources, and tools.

Intelligence After Next

MITRE strives to stimulate thought, dialogue, and action for national security leaders developing the plans, policy, and programs to guide the nation. This series of original papers is focused on the issues, policies, capabilities, and concerns of the Intelligence Community's analytical workforce as it prepares for the future. Our intent is to share our unique insights and perspectives surrounding a significant national security concern or a persistent or emerging threat, or to detail the integrated solutions and enabling technologies needed to ensure the success of the IC's analytical community in the post-COVID-19 world.

About MITRE

MITRE's mission-driven teams are dedicated to solving problems for a safer world. Through our public-private partnerships and federally funded R&D centers, we work across government and in partnership with industry to tackle challenges to the safety, stability, and well-being of our nation.

MITRE | Center for Data-Driven Policy

SEPTEMBER 2021