2022

Structured and Unstructured Data Sciences and Business Intelligence for Analyzing Requirements Post Mortem

Kendall, Walter Anthony; Zhao, Ying; Schwamm, Riqui

Monterey, California: Naval Postgraduate School

MONTEREY, CALIFORNIA

STRUCTURED AND UNSTRUCTURED DATA SCIENCES AND BUSINESS INTELLIGENCE FOR ANALYZING REQUIREMENTS POST MORTEM

EXECUTIVE SUMMARY

Principal Investigator (PI): Dr. Ying Zhao, Information Sciences (IS)

Additional Researcher(s): Mr. Tony Kendall, IS and Mr. Riqui Schwamm, IS

Student Participation: No students participated in this research project.

Prepared for:
Topic Sponsor Lead Organization: N8 - Integration of Capabilities & Resources
Topic Sponsor Name(s): CAPT Christopher Gilmore and Christopher Marsh
Topic Sponsor Contact Information: (703)6922770, christopher.s.gilmor@navy.mil and christopher.d.marsh4.ctr@us.navy.mil

This research is supported by funding from the
Naval Postgraduate School, Naval Research Program (PE 0605853N/2098).

Project Summary

The US Navy systems may have unexpected significant cost growth for many reasons. The Office of the Chief of Naval Operations (OPNAV) manually and periodically reviews big data (structured and unstructured data) that were created within the Department of Defense requirements process to identify the programs that create excessive cost or cost growth. This research explores two questions:

1. What are the common elements of requirements that create excessive cost growth in Navy systems?
2. Assuming the elements are identified, what is the risk (likelihood and magnitude) of cost growth from common elements for both procurement and sustainment costs?

We applied classic data sciences and business intelligence tools towards a more advanced artificial general intelligence framework to analyze structured and unstructured data and identify elements and factors that create excessive cost growth. We found patterns and deep causes for high cost or cost growth programs using lexical link analysis (LLA; Zhao & Zhou, 2014), natural language processing (NLP) tools, a semantic network analyzer, anomaly detection, and causal learning concepts (Pearl, 2018; Pearl & Mackenzie, 2018). Programs with anomalous characteristics can lead to high costs or high growth. These tools provide counterfactual and drill-down discovery of the key words that explain the deep causes of cost growth. The recommendations are to apply these tools for the total benefits of analyzing Navy programs and requirements of post mortem data, towards modernizing the OPNAV's Program Budget Information System (PBIS) to become a knowledge system that can effectively learn from historical data to make better risk predictions and decisions for the future Program Objectives Memorandum (POM).

Keywords: *lexical link analysis, LLA, named entity extraction, NEE, parts of speech tagging, POS, spaCy, semantic network analysis, SNA, centrality measures, unsupervised machine learning, transformers, Program Objectives Memorandum, POM, Program Budget Information System, PBIS*

Background

Navy systems may have unexpected significant cost growth for many reasons. The US Navy's OPNAV is charged, among other responsibilities, with executing the planning, programming, budgeting, and execution (PPBE) process through a series of concurrent annual planning cycles guided by a Program Objectives Memorandum (POM), collectively referred to as POM-Year X (C. Marsh, email to author, November 4, 2022).

The objective is to leverage advanced analytics to help the OPNAV understand the common elements and causes of existing Navy systems that have significant cost growth from historical data, requirements documents, and open-source media.

The PBIS has been modernized as an authoritative knowledge system including historical data of planned and executed POM information and spending each year. Data relevant to PBIS include structured data and unstructured data. For example, structured data include number of platforms procured and procurement and sustainment costs for Navy systems. Program elements (PEs) contain PPBE information as well as unstructured data of unclassified high-level program descriptions and their elements. Initial capability documents (ICDs), key performance parameters

(KPPs), or key-systems attributes (KSAs) from capability development documents (CDDs) and operational requirements documents (ORDs) are classified data sources from previous requirements processes that may have contributed to excessive cost growth. These data can be structured, such as KPPs and KSAs, and unstructured, such as PEs, ICDs, and CDDs.

We applied two categories of methods: 1. classic data sciences and business intelligence tools and 2. an artificial general intelligence framework to address the needs and research questions to analyze structured and unstructured data together and correlate them with excessive cost or cost growth of Navy systems. Specifically, we applied LLA, a semantic network analyzer, anomaly detection, and causal learning to discover patterns and deep causes that can lead to high cost or cost growth.

We analyzed two unclassified data sets provided by the topic sponsors. The first data set included seven PE documents that are processed using the LLA, artificial general intelligence NLP named entity extraction (NEE) and parts of speech (POS) tagging tools. POS features include extracted noun and verb word features. NEE features include extracted person, organization, location, product, money, event, law, language, date, time, percent, ordinal, cardinal, quantity, nationality or religious group, infrastructure, and work of art.

To discover the anomalous characteristics, we first applied LLA to compute the similarity of every two pairs of programs, then applied community finding and centrality calculation algorithms to discover the programs that are far away from community centers or on the edges of the semantic networks, which are indicators of anomalies. We used a semantic network analyzer to visualize that these Navy systems located in the center or edge of the semantic networks. The number of links are also indicators of system independences represented in the word feature networks discovered by LLA. Less linked PEs are anomalous via the unsupervised learning because they may have more unique features or innovations. We also used LLA's drill-down search capability and counterfactual reasoning of causal inferences to narrow down the key words as potential causes for the anomalous characteristics.

Findings and Conclusions

For the first data set, we found that the cost growth does not correlate with the popular, emerging, and anomalous categories of PE documents; however, some evidence shows that it may correlate with the innovativeness of the programs, which can be measured using the number of unique features of a Navy system. We also found that the numbers of people and organizations detected in the PE documents do not seem to correlate with the cost growth; however, the number of nouns (concepts) detected in the data may correlate with the cost growth.

For the second data set, we found that programs with anomalous characteristics can lead to high cost or high growth. Some anomalous PEs represented as nodes in a semantic network are less linked to other PEs and locate at the edge of the semantic network of all PEs. We also found that the number of such links computed by the unsupervised LLA are indicators of system innovations (e.g., unique features) and system independences, which are correlated with excessive cost or cost growth. We also found patterns of key words used in the PEs, i.e., key words used more than once and across multiple PEs, such as "sole source," "recurring cost," and "recurring engineering" can be the deep causes for the high cost or high cost growth.

The deliverables include the presentation and demonstration shown to the topic sponsors on November 4, 2022 and submission of a paper proposal to the 20th Annual Acquisition Research Symposium on May, 2023.

The recommendation for the sponsor to use the findings (e.g., semantic network independency and keywords) is to validate them with more PEs and build the knowledge into the risk assessment tools for new and future PEs.

Recommendations for Further Research

The recommendations for future work in this area are to

- Apply the analytic tools explored in this project to the other classified data sets such as initial capability documents, key performance parameters, key-systems attributes, capability development documents, and operational requirements documents from previous requirements processes.
- Automate the planning, programming, budgeting, and execution tasks by integrating classic data sciences, business intelligence tools, and the artificial general intelligence framework and providing capabilities of index, search, link analysis, semantic network analysis, and causality analysis to process structured and unstructured data simultaneously, allowing variety of questions of decision makers answered with more ease and better explanations.
- Scale up the combined analytic tools to the Office of the Chief of Naval Operations' Program Budget Information System (PBIS) data to predict the risk of cost growth for new and future Navy systems.
- Enable the PBIS to become a knowledge system that can effectively learn from historical data to make better assessments and decisions for the future Program Objectives Memorandum, which is consistent with the goals of the U.S. National Security Commission and industrial trends on artificial intelligence and artificial general intelligence (The National Security Commission on Artificial Intelligence, 2021; Farris, 2021).

References

Pearl, J. (2018). *The seven pillars of causal reasoning with reflections on machine learning* (R-481). Association for Computing Machinery. http://ftp.cs.ucla.edu/pub/stat_ser/r481.pdf

Pearl, J. & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic books.

Farris, T. (2021, October 27). *Eric Schmidt — The promises and perils of AI, the future of warfare, profound revolutions, and more* [Video]. YouTube. https://www.youtube.com/watch?v=AGNImy8E02w

The National Security Commission on Artificial Intelligence. (2021). The final report. https://www.nscai.gov/2021-final-report/

Zhao, Y., & Zhou, C. (2014). *System and method for knowledge pattern search from networked agents* (U.S. Patent No. 8,903,756). U.S. Patent and Trademark Office. https://www.google.com/patents/US8903756

Acronyms

AI	artificial intelligence
CDDs	capability development documents
ICDs	initial capability documents
KPPs	key performance parameters
KSAs	key-systems attributes
LLA	lexical link analysis
ML	machine learning
NLP	natural language processing
OPNAV	Office of the Chief of Naval Operations
ORDs	operational requirements documents
PBIS	Program Budget Information System
PE	program elements
POM	Program Objectives Memorandum
PPBE	planning, programming, budgeting, and execution
SNA	semantic network analyzer