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

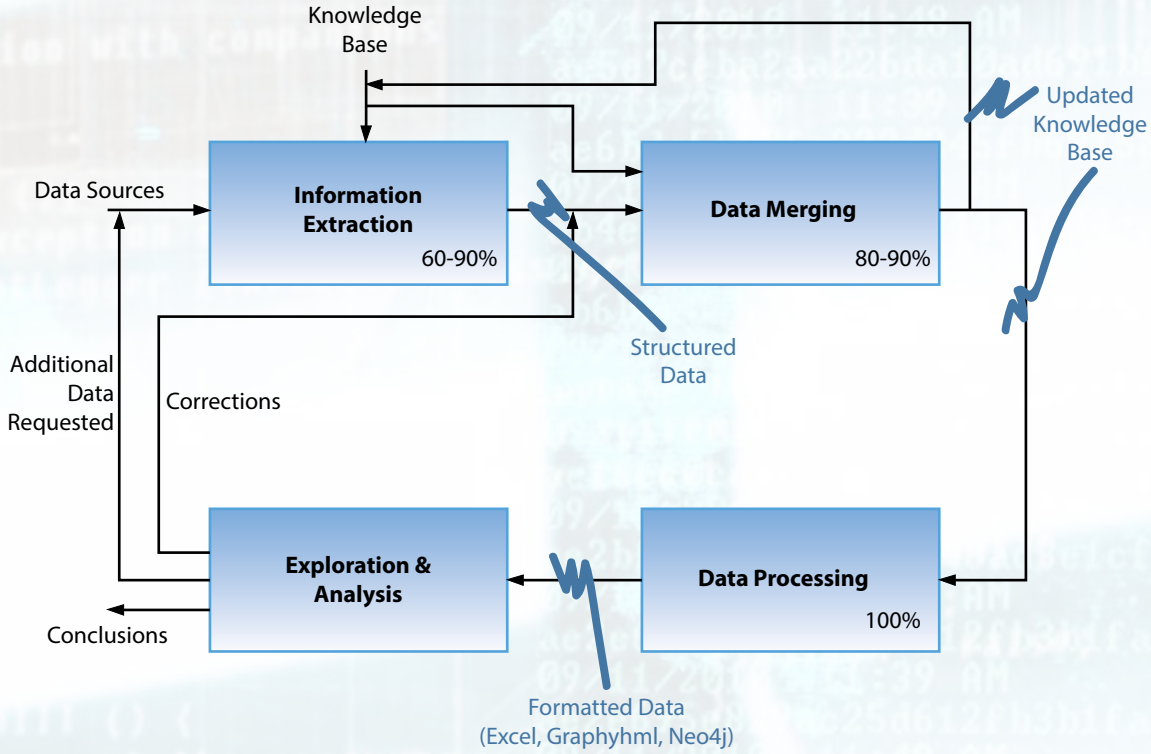

Figure 1. Process Diagram

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.

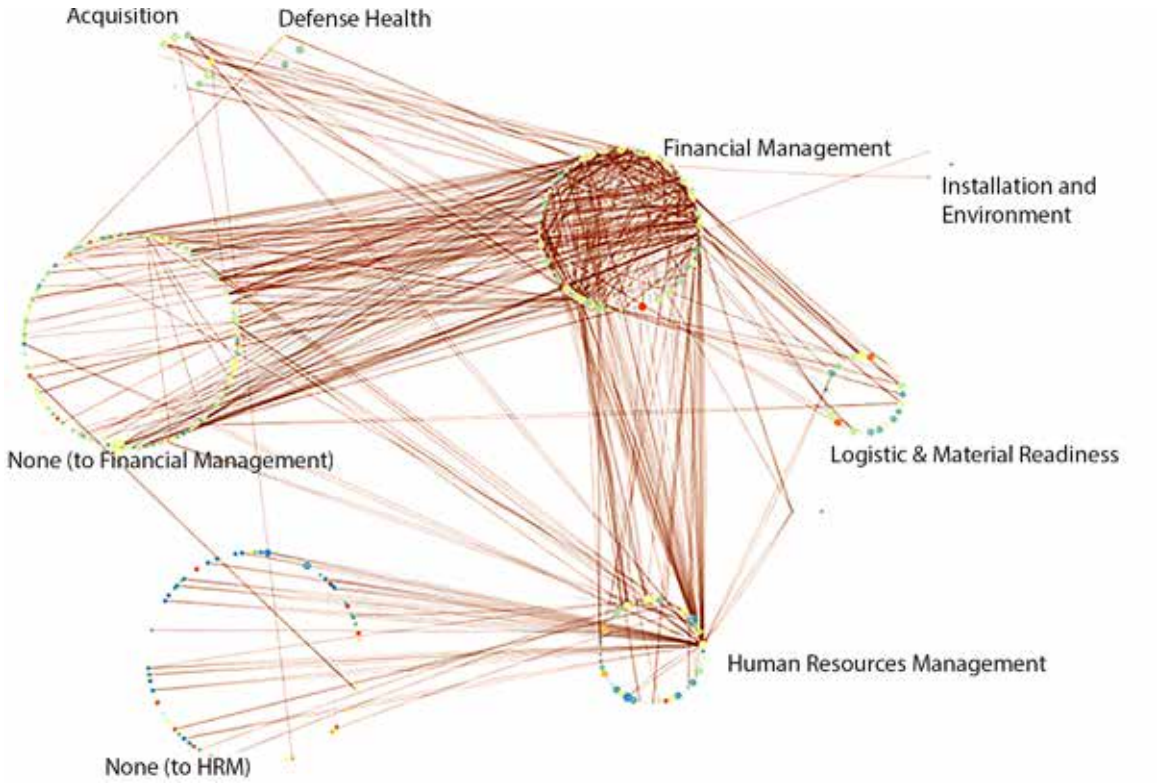

Figure 2. Group Attributes Layout by Function

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.