Case Study

Automating Data Pipelines

To Identify Meaningful Connections Between Network Entities

Elder Research developed an automated data pipeline to cleanse data and feed a data visualization tool used to identify and explore document preparer network relationships. The solution enabled the client to automate significant portions of work, make data-driven decisions, prioritize resources, and gain new business value from the data.

INDUSTRY

- Government

BUSINESS NEED

- Analyze millions of documents to correct erroneous and missing data and enable visualization, identification, and exploration of document preparer networks

SOLUTION

- Applied advanced analytics and data visualization to identify meaningful connections between network entities

BENEFIT

- Automated 40% of the cases investigated for improper preparer identification
- Provided a new way for the client to understand the preparer population
- Enabled the client to make data-driven decisions, prioritize resources, and gain new business value from the data

The Challenge

Elder Research was tasked to identify networks of document preparers who worked together in a given year. Documents with particular information in common indicated a possible network connection. The goals were to enhance the client's capabilities in two main ways:

1. Improve efficiency of finding documents with incorrect preparer identification and re-assigning them to the correct preparer identification when possible.
2. Enable the analysts to consider preparer networks instead of only individual preparers to better deploy investigative resources.

The Solution

The project required several interrelated stages of data analysis using multiple data sources and formats. Since it was common for the documents to have typographical and other errors – including potentially intentional misidentification of the preparer -- preparers could appear to be linked when they did not actually work together. Elder Research used extensive data validation procedures to account for missing data and ensure that the documents identified the proper preparer.

Next, we designed an entity resolution process to compare information on different documents and identify pairs with a statistically strong match. Entity resolution resolves multiple labels for individuals, in this case document preparers, into a single resolved entity that can be analyzed for relationships. The example below shows two records that could be the same entity.

Name	Address	City	State	Phone
Starbucks	3457 Hillsborough Road	Durham	NC	NULL
Starbacks	Hillsborough Rd	Durham	NULL	919-333-4444

Because of the many ways that documents could err on preparer information a modular process was implemented, handling even challenging cases such as where initials were used instead of full names. Entity resolution improved network identification accuracy and automated some time-consuming tasks previously performed manually by the client. Making pairwise comparisons between millions of records is computationally intensive so Elder Research devised a data compression

Headquarters
300 W. Main Street, Suite 301
Charlottesville, VA 22903
(434) 973-7673
www.elderresearch.com

Elder Research
Data Science & Predictive Analytics

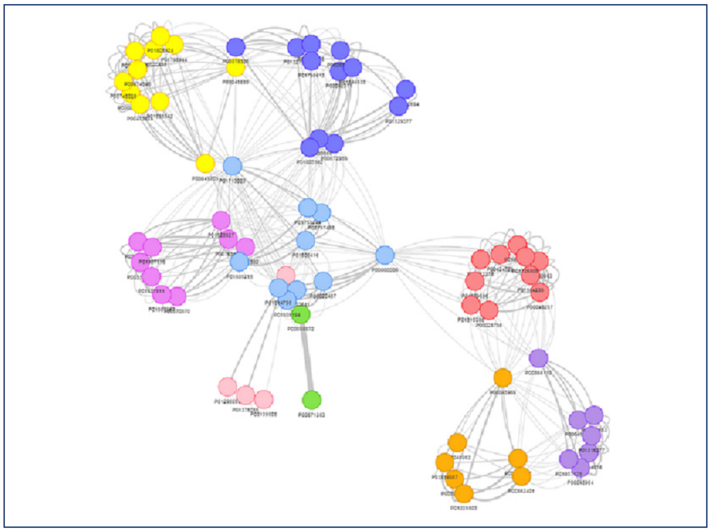
Office Locations
Arlington, VA
Linthicum, MD
Raleigh, NC

strategy that transformed 85 million documents into 2 million document groups. This solution retained the relevant information and solved the computational challenge.

For the preliminary phase of network identification Elder Research quantified the strength of relationship between each pair of document preparers. This strength was based on the volume of the preparers' documents that had information in common, and how rare those pieces of information were in the preparer population. The idea behind this technique came via Elder Research's extensive experience with text mining, as exemplified by the award-winning book, Practical Text Mining, co-authored by senior Elder Research scientists Andrew Fast and John Elder and four others (Elsevier, 2012).

Due to the nature of the data the resulting networks were impractically large, with multiple degrees of linkage between preparers. This required further phases of network analysis techniques to sort through the web of connections and boil down the links to the most likely networks of preparers who worked together, as shown in the example below.

To make the graphs actionable on live data, automated scripts were written to create a data processing pipeline for raw data to feed a graph database. The graph database enabled the client to interactively explore and visualize relationships among preparers. Elder Research then helped the client create a second, simplified graph database along with query templates to allow analysts to also explore preparer connections based on the raw document data (without the pre-processing analytics). Further, to enable end users lacking programming skills to search and explore preparer relationships, Elder Research deployed its proprietary browser-based network visualization tool.



Network of Preparers Visualization

Results

The solution deployed by Elder Research enabled the client to make data-driven decisions and prioritize scarce investigative resources. The solution streamlined 40% of the improper identification cases investigated by automatically suggesting likely matches —reducing a manual process that took 20 minutes per case to one taking less than a minute. In another 45% of cases, the solution partially automated the investigative process by linking questionable preparers to a particular network so their true identity could be determined.

In addition to providing significant time savings, the graph databases helped the client to gain new business value by analyzing the data in novel ways. End users can investigate multiple risk metrics and characteristics of interest at a network level, rather than only at the individual preparer level. They are able to see how preparers are related, and how closely. By writing graph database queries, analysts can answer their internal clients' business questions, and they can make some changes to the database and algorithms to meet the evolving needs of their organization.

About Elder Research

Elder Research is a recognized leader in the science, practice, and technology of advanced analytics. We have helped government agencies and Fortune Global 500® companies solve real-world problems across diverse industries. Our areas of expertise include data science, text mining, data visualization, scientific software engineering, and technical teaching. With experience in diverse projects and algorithms, advanced validation techniques, and innovative model combination methods (ensembles), Elder Research can maximize project success to ensure a continued return on analytics investment.

Headquarters
300 W. Main Street, Suite 301
Charlottesville, VA 22903
(434) 973-7673
www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

Office Locations
Arlington, VA
Linthicum, MD
Raleigh, NC