Case Study

AUTOMATING TEXTUAL DATA DISCOVERY AND ANALYSIS

TO TRACK AND EVALUATE KEY EVENTS RELATED TO INFECTIOUS ANIMAL DISEASES

Elder Research designed and deployed tools for high-end text mining to search, index, and automatically classify information related to animal infectious diseases.

INDUSTRY

» Defense and intelligence

BUSINESS NEED

» Required a solution that would enable analysts to more easily track and evaluate key events related to animal diseases for any country

SOLUTION

» Designed a data mining and visualization system to search, index, and automatically classify information related to significant animal infectious diseases

» Provided automated alerts and event reports

BENEFIT

» Enabled analysts to validate incidents of animal disease and make recommendations for dealing with them at the earliest stage possible

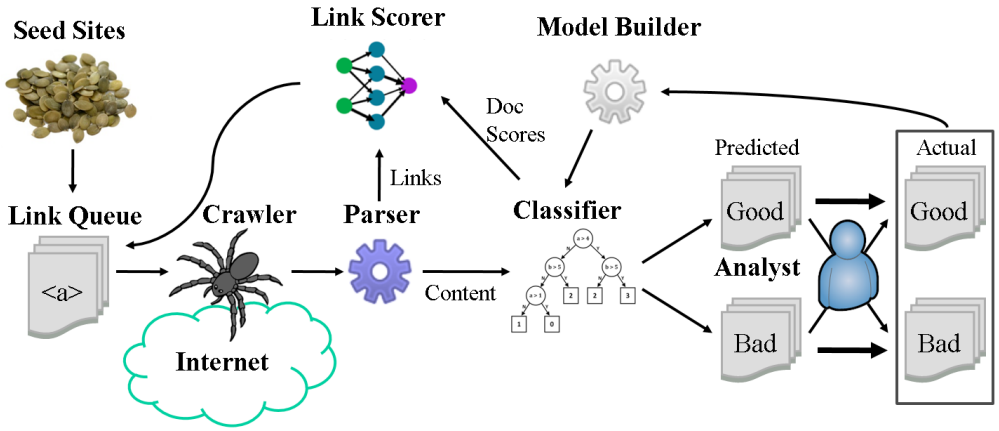
The Challenge

This project supported a federal agency tasked with coordinating interdisciplinary activities focused on protecting America's agricultural infrastructure and economy from endemic and emerging biological threats. Tracking and responding to high consequence infectious animal disease events both domestically and abroad is critical for national security and economic stability. The goal was to enhance agency analyses through real-time, state-of-the-art tools that could acquire, assess, and integrate comprehensive data from the internet and commercial, government, and NGO databases.

The Solution

Elder Research combined state-of-the-art data mining and machine learning tools with best practices for data integration to enhance analysts' efficiency in developing country-specific veterinary capability studies. A visual representation of the system is shown below.

The proprietary solution employed cutting-edge "learned rule" natural language processing, which did not rely on dictionaries and heuristics. All of the clustering and "document fingerprinting" used robust rules learned from observed outcomes. Available technologies and software tools were evaluated to determine the most cost-effective solution for the real-time data streaming architecture and



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

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

Satellite Locations
Arlington, VA
Linthicum, MD
Raleigh, NC

the best way to provide access to the results. Analytical models were implemented that were compatible with the client's existing systems and workflows, and a user-friendly, real-time web-based search interface provided analysts easy access to the results.

The strength of this system was its power and simplicity. The flexible, statistical model used the full-text of documents rather than just keywords to provide context for the desired search topics. Using a set of training documents supplied by the analyst, the focused web crawler used data mining technology to locate similar relevant documents and store them on the capabilities database. It also actively monitored the changing data sources for incoming documents matching the search profile. By configuring automated alerts, analysts could receive an electronic mail or a text message when new documents of interest arrived. Intuitive tagging and scoring features enabled analysts to supply feedback on documents to update the model and improve the focus of the system — enabling the system to truly learn.

That is, after an analyst supplies a judgment of "thumbs up or down" on a particular document, the system gains a labeled training example of what the analyst is, or isn't, looking for. This database of labeled cases is fuel to a machine learning model builder, which updates a model, then can rapidly score all new candidate documents. Those documents with the highest ratings are presented to the analyst as the most likely documents of interest. Thus, each judgment from the analyst can almost immediately re-order the vast list of candidate documents — greatly enhancing analyst efficiency.

Based on input from the client the system was designed to take advantage of analysts' expertise in many different forms and employ multiple methods for gathering information:

- External Search Engine Module – Ran a parallel country-specific keyword search against content already stored in the database and three major search engines. The keyword search allowed analysts to restrict searches based on document tags and ratings, or specific web addresses.
- Search Monitor Module – Ran and saved sets of keyword searches automatically, and tagged and stored that content in the database for future searching. The searches could be set to repeat at a defined time intervals.
- Site Monitor Module – Monitored or "scraped" an entire domain or website and added that content into the database for future retrieval. This feature enabled a more in-depth search of known valuable sites compared to public search engines that may only skim the surface of the site. The site monitor could be set to automatically refresh.
- Import Modules – Imported documents from a hard drive, email or FTP directory.
- RSS Reader – Imported RSS feeds of valuable content.

Results

Automating the process of discovery and synthesis of textual data from current and reliable sources made the work of sifting through huge databases and incoming information streams manageable, intuitive and efficient, and provided reliable event reports to enable analysts to validate incidents of animal disease and make recommendations for dealing with them at the earliest stage possible.

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

Satellite Locations
Arlington, VA
Linthicum, MD
Raleigh, NC