VAST 2007 Contest

Interactive Poster: Data Analysis Using NdCore and REGGAE

Lynn Schwendiman, Jonathan McLean, Jonathan Larson

ATS Intelligent Discovery

ABSTRACT

ATS Intelligent Discovery analyzed the VAST 2007 contest data set using two of its proprietary applications, NdCore and REGGAE (Relationship Generating Graph Analysis Engine). The paper describes these tools and how they were used to discover the contest's scenarios of wildlife law enforcement, endangered species issues, and ecoterrorism.

CR Categories and Subject Descriptors: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces; I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques

Additional Keywords: visualization, visual analytics, text analysis, data discovery

1 INTRODUCTION

NdCore and REGGAE (Relationship Generating Graph Analysis Engine) are two proprietary data analysis and discovery applications developed by ATS. NdCore is our current production analytical tool. REGGAE is a prototype of our patent-pending next generation data analytics engine, currently under development.

1.1 NdCore

NdCore is a powerful tool for integrating and analyzing large volumes of data from disparate sources. NdCore ingests both structured and unstructured data from relational databases and a variety of file formats, including plain text, MSWord, PDF, and XML, in any combination, into a single repository for analysis.

NdCore's text analysis generates two- and three-word concepts based on the frequency, usage, and relative proximity of words within the ingested documents. The NdCore Concept Builder guides the user in defining a multi-term search phrase based on the frequency of word combination occurrences. Concept Builder uses actual document text to suggest the most common word combinations. It can also suggest alternate terms (words used in a similar context, stems, inflections, sound-alikes) to include in the search. By suggesting words as the search phrase is created, Concept Builder helps the user make intelligent decisions about which words and combinations will help find documents of interest. The web-based user interface allows the user to quickly browse through the selected documents and to find other documents similar to them.

ATS Intelligent Discovery
3505 NW Anderson Hill Road, Suite 200
Silverdale, WA 98383
lynn.schwendiman@atsid.com
jonathan.mclean@atsid.com
jonathan.larson@atsid.com

1.2 REGGAE

REGGAE, a prototype currently under development, provides the associative search capabilities of a graph database and the tabular capabilities of an RDBMS while also incorporating the multidimensional analysis seen in a multidimensional database.

REGGAE is specifically tailored for entity relationship generation and analytics. It utilizes a novel two-tiered context-based graph architecture. One tier, the data layer, is populated with entity nodes. The second tier, the context layer, is populated with context nodes that represent links or relationships between entities. Duplicate entities are never inserted into the database. Instead, new relationships are generated in the context layer between the existing entities and new entities as new data is imported.

REGGAE is built on current commercial relational databases and easily integrates with existing data stores. REGGAE also includes integrated text searching and clustering capabilities for document analysis.

2 ANALYSIS APPROACH

NdCore and REGGAE were each used independently to analyze the VAST data set, NdCore for the raw text alone, REGGAE for both raw text and the preprocessed extracted entities.

2.1 Using NdCore to Analyze Raw Text

We began by importing the VAST contest document set into NdCore and performing a text analysis, which parsed the document text, built concepts, and identified related words, stems, and sound-alike words.

The general analysis approach was to use NdCore's Concept Builder to suggest combinations of search terms, perform the search, and read the returned documents, noting people, organizations, locations, activities, and events to use in subsequent iterative searches. Once a document of interest was identified, NdCore's "find similar documents" feature was extremely helpful in discovering additional relevant documents.

For example, starting with the search terms "endangered species", Concept Builder suggested a third term, "CITES", to complete the concept. The search returned five documents, two of which concerned the Assan Circus and its wildlife smuggling activities. "Find similar" performed on the Assan documents turned up a third relevant document. Thus we were able to discern the likely involvement of the Assan Circus in the contest scenario after examining just six documents.

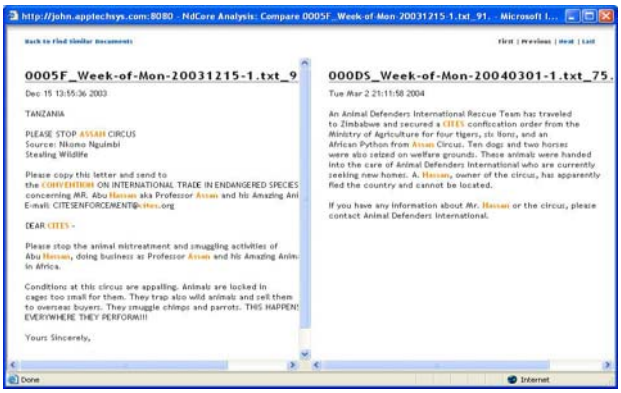

Figure 1. NdCore's similar document comparison

2.2 Using REGGAE to Analyze Structured Data

REGGAE is geared primarily toward analysis of discrete, structured data but also includes an integrated text searching capability, so both the VAST entity extraction data and the raw text were imported and processed.

For the analysis in REGGAE, we used two kinds of queries: keyword searches on document text and single cell queries on the structured data. Single cell queries return nodes which are placed on a chart in the structured data viewer. Keyword searches return a set of documents to the document viewer. Documents of interest can be selected and exported to the structured data viewer where they appear as nodes on a chart.

Any node on the chart can be expanded to reveal all of its associated structured data and the connections among them. In the case of the VAST data, expansion of a document node reveals its extracted entities and expansion of an extracted entity node reveals its associated documents. The full text associated with a document node can be accessed directly from the chart.

The general analysis approach for REGGAE was to run a query, either keyword or single cell, place returned nodes of interest on a chart, and start expanding nodes (generally documents, people, and organizations). When the chart became crowded, we constrained it to exclude nodes that had fewer than two connections. The people and organizations remaining on the chart were candidates for further investigation. We also used REGGAE's "Find Similar" feature to suggest new searches and to guide further expansion of the chart. Once an entity of interest was identified, similar entities could be found based on their common connections.

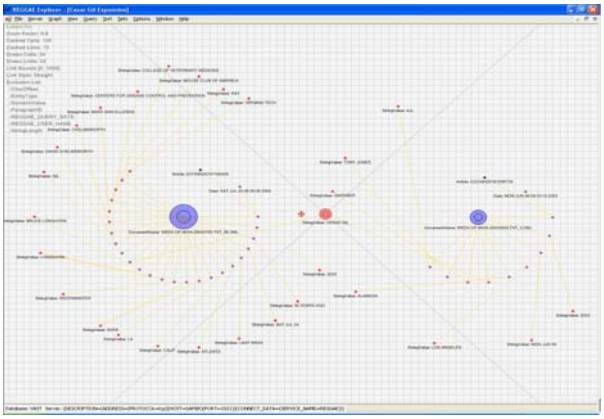

Figure 2. REGGAE chart with two expanded document nodes

3 Strengths and Limitations

3.1 NdCore

NdCore allows analysts to easily load their data for analysis. Once loaded, it provides a discovery workflow that guides an analyst's search through textual data. NdCore also provides a powerful query expansion utility that allows analysts to search on topics to detect misspellings and alternative term usages. Once a document of interest has been located, NdCore also provides excellent facilities for locating other documents of interest and managing the analyst's workflow. Summarizations of each document are provided as well as tracking data that shows the analyst which documents have already been reviewed.

Currently, NdCore needs more visual analytics development. While the analytics capabilities of it are very powerful, viewing and collecting the results of an analysis proved to be a challenge.

3.2 REGGAE

REGGAE's greatest strengths lie in its unique engine architecture and analytic algorithms. It has powerful similarity searching capabilities that were leveraged on the entity extracted data. It also has an integrated environment that is useful for browsing connections across all of the structured data.

Currently, REGGAE is still in its prototype stage and thus lacks a polished user interaction. While it performs powerful data mining queries, it lacks a solid analyst workflow.

4 Lessons Learned from the Contest

The VAST contest was a great opportunity to exercise our analytics tools. Working with a realistic data set to uncover a plausible scenario was enlightening. The experience left us with these thoughts:

- Collaboration is key. Analytics tools can take you only so far. At some point you have to make an imaginative leap to fill in the missing details and formulate a plausible hypothesis. Having more than one analyst reviewing and discussing the data and proposing theories is highly desirable. This may seem obvious, but participation in the contest drove the point home.
- The analyst's workflow is an important consideration in designing the tool's interactions. An integrated component for collecting and organizing analysis results and ideas as they occur would be extremely helpful.
- It's all too easy to neglect the data that your analytics tools don't handle. In our case, the jpg images were given less attention than they deserved. We were so caught up in using our tools to analyze the text documents that we failed to connect Mercurio Navarro (who appears in the tropical fish importers spreadsheet, but in none of the text documents) with the initials "M. N." that appeared in one of the jpg images. This turned out to be a vital link that we missed completely. The broader array of data that your analytics tools can accommodate, the better.

5 Conclusions

NdCore and REGGAE were used successfully to uncover most of the details of the VAST contest scenarios. NdCore's powerful text analysis capabilities are complemented by REGGAE and its visual analytics. REGGAE is in the early stages of development but has the potential of becoming a powerful analytics tool.