Advanced Methods Allow Processing and Searching of Very Large Document Collections

Contacts: Daniel Dunlavy, Computer Science & Informatics, *dmdunla@sandia.gov*
Tim Shead, Data Analysis and Visualization, *tshead@sandia.gov*
Patricia Crossno, Data Analysis and Visualization, *pjcross@sandia.gov*

…we have developed a flexible suite of scalable text analysis software components and applications, visual algorithm analysis tools, and a software system for creating and evaluating ensemble models for general machine learning tasks. These tools are currently being used to solve problems in the areas of nonproliferation, technology surprise, nuclear materials attribution and computer simulation data analysis.

Intelligence analysts have a big data problem. They answer questions of national security under extreme time pressure, but current tools do not scale to the volume of data they must consider. No end-to-end scalable visual text analysis capabilities exist today, and this prevents analysts from exploring, analyzing, and annotating existing petascale document collections. For this project, we have developed a flexible suite of scalable text analysis software components and applications, visual algorithm analysis tools, and a software system for creating and evaluating ensemble models for general machine learning tasks. These tools are currently being used to solve problems in the areas of nonproliferation, technology surprise, nuclear materials attribution and computer simulation data analysis.

The main contributions to date of this project include the design and development of the ParaText™ Server (Figure 1), Algebraic Engine, LSAView (Figure 2), and HEMLOCK data modeling and analysis software frameworks and applications. The ParaText™ Server is both a platform for development and experimentation of large-scale text analysis algorithms. It provides a web services API and includes components for text processing and latent semantic analysis (i.e., statistical language modeling, conceptual information retrieval, and similarity analysis based on term usage). Figure 1 presents the application stack and an HTML example of the ParaText™ Server.

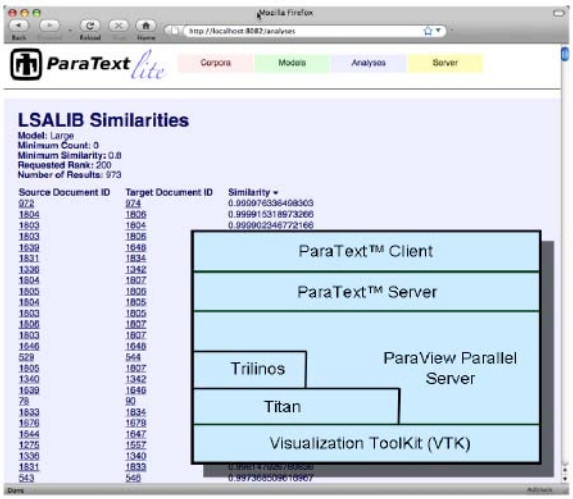

Figure 1. The ParaText™ Server is built on existing large-scale analysis capabilities developed at Sandia (inset). Through its web services API, ParaText™ Server can deliver text analysis capabilities to any desktop with a web browser.

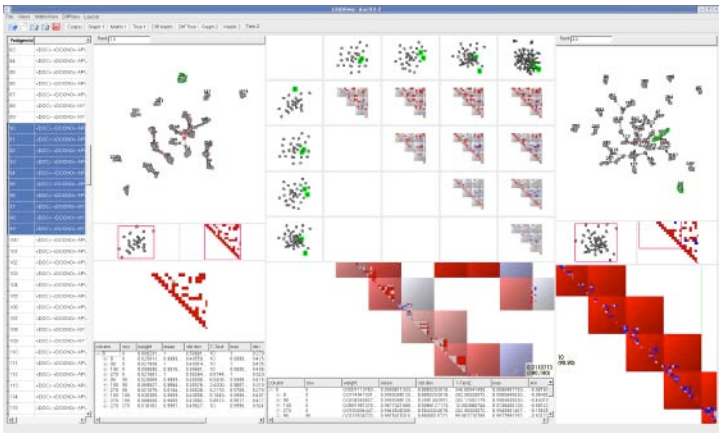

Figure 2. The LSAView application is being used to visually analyze text analysis algorithms and their impact on end-user interpretation of results.

The Algebraic Engine generalizes the capabilities of the ParaText Server to provide an architecture for the storage and processing of algebraic data analysis models. It consists of a set of Visualization Toolkit (VTK) components for creating and processing algebraic models (e.g., matrix SVD and tensor PARAFAC models), analyzing data using such models (e.g., similarities between data objects), and interacting with SQL databases for model caching. The Algebraic Engine is currently being used in the following applications:

- ParaText™: scalable text analysis
- OverView: open source information visualization framework
- TreeView: temporal information visualization for nuclear materials forensic attribution

- LSAView: text algorithm analysis
- LDRDView: funding portfolio analysis
- ParaSpace: simulation data analysis

LSAView is a visualization application for analyzing and exploring the impact of informatics algorithms on end-user visual analysis of data. LSAView aids developers in the process of algorithm design and aids users in the discovery of optimal parameter settings for specific data analysis problems. Figure 2 illustrates the current features:

- Side-by-side data comparison views for two sets of algorithm parameters
- Small multiple view for analyzing three or more parameter sets simultaneously
- Linked document, graph, matrix, and tree data views
- Interactive, zoomable, hierarchical sparse matrix and difference views
- Statistical inference tests used in highlighting novel parameter impact

For this project, we have also been investigating machine learning methods for incorporating relevance feedback into models and automatically tuning data analyses to align with end-user requirements. To this end, we have developed the Heterogeneous Ensemble Machine Learning Open Classification Kit (HEMLOCK) [1] for creating and evaluating ensemble models aimed at classification problems. More specifically, HEMLOCK can generate metamodels of classifier models of different types (e.g., probabilistic naïve Bayes, support vector machine, and decision tree models). Several methods for combining models (e.g., majority voting, sum rule, stacking, etc.) and creating diverse models (e.g, bagging, random forests, random subspace method, etc.) are currently available and have been investigated in solving classification problems using images. Figure 3 illustrates the performance of HEMLOCK ensemble classifiers over individual and homogeneous ensemble classifiers. By combining diverse individual models to form ensembles of models, HEMLOCK models classify new data more effectively.

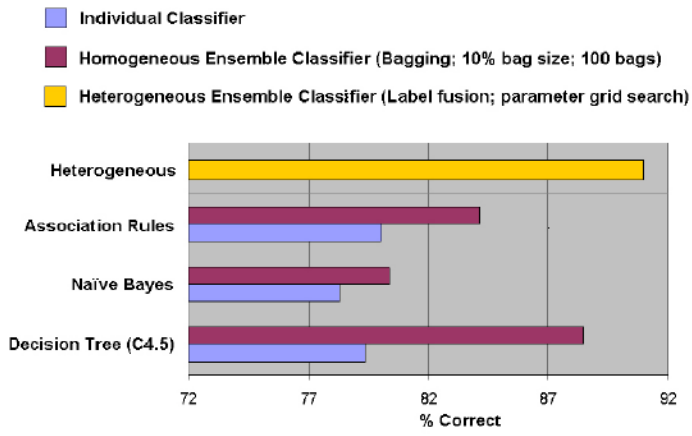

Figure 3. Results using HEMLOCK ensembles (yellow) indicate improved performance over existing individual and homogeneous ensemble models on a standard image classification task.

Reference

1. Gilpin S. A., and Dunlavy, D. M. (2008). Heterogeneous Ensemble Classification, *Computer Science Research Institute Summer Proceedings*, SAND 2008-8257P, Sandia National Laboratories, Albuquerque, NM.

This work has been supported by SNL's Laboratory Directed Research and Development (LDRD) Program.