ParaText™: Scalable Solutions for Processing and Searching Very Large Document Collections

Sandia National Laboratories

Daniel M. Dunlavy, Heidi Ammerlahn, Timothy M. Shead, Patricia J. Crossno, Peter A. Chew, Sean A. Gilpin

Problem

Big Data Problems

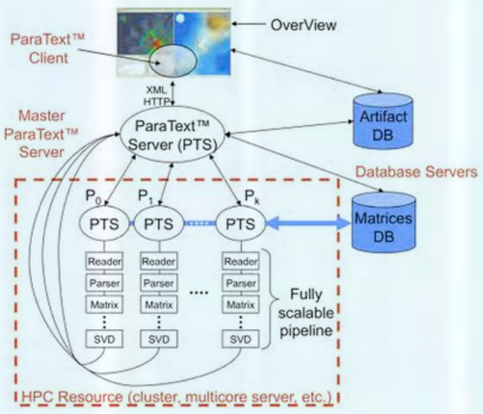
Intelligence analysts and cyber security experts have big data problems, answering questions of national security under extreme time pressure. However, current tools do not scale to the volume of data (news articles, reports, network traffic payloads, etc.) they must consider. Determining relationships between documents as well as between entities (e.g., people, places, organizations, etc.) and terms within those documents is often a crucial task required for answering important security questions.

Approach

R&D Methodology: Overview

- Scalable text analysis framework for analyzing conceptual relationships between documents, between terms (including named entities) in documents, and between terms and documents.
- Visual analysis of graph-based relational informatics algorithms for determining optimal parameters as they impact decision-making applications.
- Heterogeneous ensemble data classification modeling for improving machine learning models with little or no input from users. Learning models can be used for providing feedback to text analysis algorithms to incorporate analyst subject matter expertise.
- Unsupervised natural language modeling based on statistical methods require little or no manually annotated training data for extracting parts of speech, named entities (persons, locations, organizations, etc.), significant phrases, etc.

R&D Methodology: ParaText™



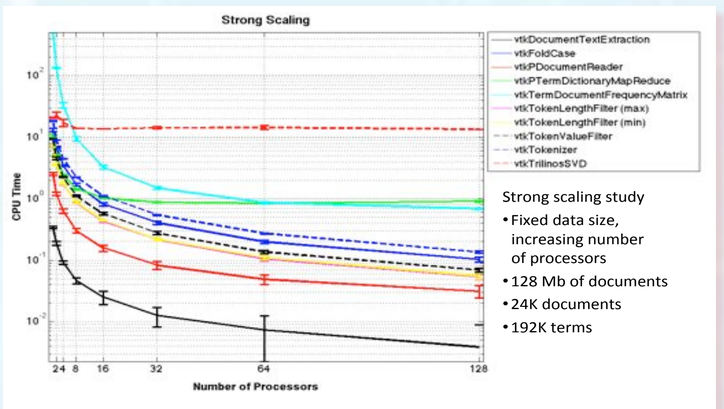
- End-to-end scalable system for statistical text analysis and visualization
- Capabilities: latent semantic analysis (LSA), n-gram analysis, unsupervised part of speech and named entity extraction
- Leveraging existing Sandia tools: Titan, ParaView, Trilinos, LSALIB, MapReduce
- Client-server, web service, and stand-alone versions

Results

Ensemble Machine Learning

- Handwritten digit classification example
- Heterogeneous ensembles outperform homogeneous ensembles and individual classifier models

ParaText™ Scalability

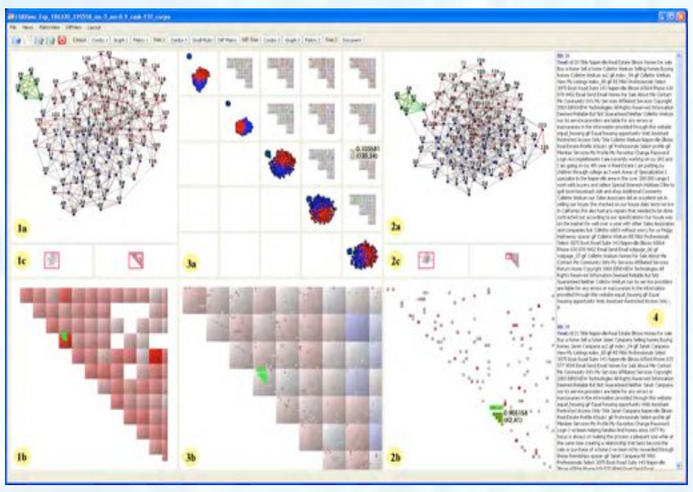


Strong scaling study
- Fixed data size, increasing number of processors
- 128 Mb of documents
- 24K documents
- 192K terms

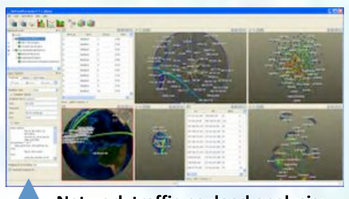
Results

Visual Algorithm Analysis

- Identification of useful choices of SVD rank and singular value rescaling in graph clustering algorithms using LSA illustratesthe utility of visual algorithm analysis over existing methods for parameter identification in data analysis.

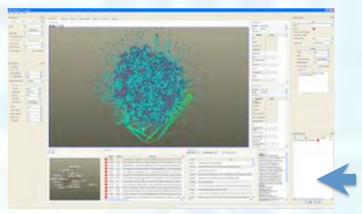


Applications

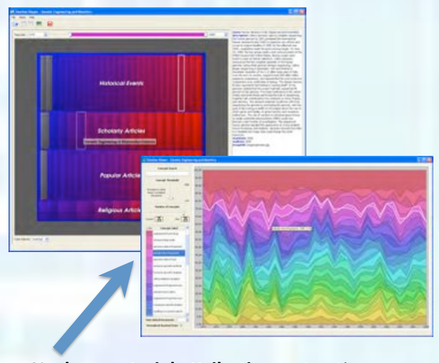


Network traffic payload analysis: conceptual document matching and clustering enables assessment of emerging cyber threats.

Nuclear nonproliferation: combining ParaText™ text analysis with graph algorithms and visualization methods provides intelligence analysts with a broad range of document analysis capabilities for assessing potential issues in the area of nonproliferation.



Technology surprise: combining scalable ParaText™ algorithms with visual relational tools aids analysts in determining technology capabilities using more data over larger time periods.



Nuclear material attribution: extracting meaningful patterns and trends as well as providing over specific time periods in nuclear-materials production information enables rapid interpretation of nuclear-forensic signatures.

Funding portfolio analysis: conceptual search of and matching between documents using one or more fields structured data (i.e., databases) helps funding managers better understand the core areas of a portfolio in the presence of diverse uses of terminology.

Significance

The ParaText™ LDRD project is impacting several critical applications of importance to Sandia, NNSA, and the nation as a whole in the areas of intelligence analysis, cyber security, and ASC modeling and simulation.

- Network traffic text payload analysis
- Nuclear material attribution
- Nuclear nonproliferation
- Technology surprise
- Funding portfolio management

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2009-4430P

National Nuclear Security Administration

Sandia National Laboratories