SANDIA REPORT
SAND2010-6269
Unlimited Release
Printed September 2010

ParaText – Scalable Solutions for Processing and Searching Very Large Document Collections: Final LDRD Report

Daniel M. Dunlavy, Timothy M. Shead, Patricia J. Crossno, Eric T. Stanton

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation,
a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's
National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online order:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2010-6269
Unlimited Release
Printed September 2010

ParaText – Scalable Solutions for Processing and Searching Very Large Document Collections: Final LDRD Report

Daniel M. Dunlavy
Computer Science and Informatics
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-MS1318

Timothy M. Shead, Patricia J. Crossno and Eric T. Stanton
Data Analysis and Visualization
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-MS1323

Abstract

This report is a summary of the accomplishments of the "Scalable Solutions for Processing and Searching Very Large Document Collections" LDRD, which ran from FY08 through FY10. Our goal was to investigate scalable text analysis; specifically, methods for information retrieval and visualization that could scale to extremely large document collections. Towards that end, we designed, implemented, and demonstrated a scalable framework for text analysis – ParaText - as a major project deliverable. Further, we demonstrated the benefits of using visual analysis in text analysis algorithm development, improved performance of heterogeneous ensemble models in data classification problems, and the advantages of information theoretic methods in user analysis and interpretation in cross language information retrieval. The project involved 5 members of the technical staff and 3 summer interns (including one who worked two summers). It resulted in a total of 14 publications, 3 new software libraries (2 open source and 1 internal to Sandia), several new end-user software applications, and over 20 presentations. Several follow-on projects have already begun or will start in FY11, with additional projects currently in proposal.

ACKNOWLEDGMENTS

This work was fully supported by Sandia’s Laboratory Directed Research & Development (LDRD) program.

We would like to thank all of the members of the ParaText LDRD project who contributed time, ideas, research, and software over the course of the project:

Daniel M. Dunlavy, PI
Heidi Ammerlahn, PM
Patricia J. Crossno
Timothy M. Shead
Eric T. Stanton
Peter A. Chew
Sean A. Gilpin (student)
Taylor P. Turpen (student)
Becca Simon (student)

We would also like to thank the following people for their suggestions, comments, and assistance throughout the ParaText project: Brian Wylie, Jason Shepherd, Andy Wilson, Nathan Fabian, Laura McNamara, Nabeel Rahal, Philip Kegelmeyer, and Shawn Martin.

CONTENTS

1. Introduction 7

2. Project Description 8
   - 2.1 Problem Description 8
   - 2.2. ParaText 9
     - 2.2.1 The ParaText Pipeline 10
     - 2.2.2 The ParaText Command Line Tools 13
     - 2.2.3 ParaText Server 13
   - 2.3. LSAView 14
   - 2.4. TextView 14
   - 2.5. HEMLOCK 17

3. Impact 19
   - 3.1 Technical Impact 19
   - 3.2 Project and Application Impact 19
     - 3.2.1 Networks Grand Challenge (NGC) 19
     - 3.2.2 ThreatView 20
     - 3.2.3 Nuclear Attribution 20
     - 3.2.4 ParaSpace 21
     - 3.2.5 LDRDView 22
     - 3.2.6 Multi-role Experiential Learning 22
   - 3.3 Programmatic Impact 22
     - 3.3.1 Funding 23
       - 3.3.1.1 CSRF/CSSE 23
       - 3.3.1.2 DOE Office of Science 23
       - 3.3.1.3 LDRD 24
       - 3.3.1.4 NMSBA 25
     - 3.3.2 Leadership and Service 25
     - 3.3.3 Workshops 27
       - 3.3.3.1 WDMDA 27
       - 3.3.3.2 VizMining 27
     - 3.3.4 Awards 27
     - 3.3.5 Students 28
       - 3.3.5.1 Sean Gilpin 28
       - 3.3.5.2 Taylor (Tad) Turpen 28
       - 3.3.5.3 Becca Simon 28
   - 3.4 Publications and Presentations 29

4. References 32

Distribution 35

FIGURES

Figure 1: Typical ParaText Pipeline 10
Figure 2. Examples of different views in the LSAView application: (1a) and (2a) Graph View, (1b) and (2b) Matrix View, (1c) and (2c) You Are Here View (3a) Small Multiples View, (3b) Difference Matrix View, and (4) Document View. 14
Figure 3: Term/Concept tab views with LSA concepts in light-blue and LDA topics in light-orange. The bipartite graph displays weighted similarities between LSA concepts on the left and LDA topics on the right. Importance-ordered term lists for concepts and topics are in the term/topic table on the right. 15
Figure 4: Document/Concept tab views: document similarity graphs (LSA on the left, LDA on the right), You Are Here views below each graph providing context, a table of document weights for each concept/topic, and document text highlighting the selected term in the term/topic table. Selected documents are shown in black in the graphs, highlighted in the table, and their text displayed. 16
Figure 5. Performance comparisons between base classifiers, homogeneous ensembles, and heterogeneous ensembles for the task of image classification... 18
Figure 6: Views of the Timeline Treemap Browser displaying the evolution of genetic engineering data set. The back image shows the treemap view of timeline elements. The front image is the trend-based view showing thematic changes over time. 20
Figure 7: HelioView application showing CCA correlations for car data set. 21
Figure 8. The LDRDView application for analyzing funding portfolios. 22

TABLES

Table 1. Leadership and service associated with the ParaText LDRD project............................ 26
Table 2. Awards associated with the ParaText LDRD project. 27

1. INTRODUCTION

The focus of the “Scalable Solutions for Processing and Searching Very Large Document Collections” LDRD was to investigate and develop scalable methods for a complete, end-to-end text modeling and analysis process, from extracting raw document information to data modeling, data analysis, and visualization. The motivation for this work included 1) the importance of text analysis in national security applications and 2) the lack of an interoperable set of scalable components for text analysis and visualization. Sandia’s expertise in data and graph analysis, matrix methods, visualization and high-performance computing made it natural for us to pursue this line of inquiry.

Our goal was to develop such a system, which in turn could be used in a variety of applications to support text modeling and analysis of extremely large document collections. Thus, we developed ParaText, which is now available through the open source Titan toolkit as part of its text analysis library. ParaText has also been integrated into several operational and research prototype applications deployed throughout Sandia and to several government customers.

The remainder of this report is a discussion of the accomplishments of the ParaText project, including its impact on current and future applications and related areas of research.

2. PROJECT DESCRIPTION

In this section, we present a description of the ParaText LDRD project. After a description of the problem addressed by the project, the organization of this section reflects the major ideas and capabilities in terms of the software capabilities developed.

2.1 Problem Description

Intelligence analysts have a big data problem. They answer questions of national security under extreme time pressure. In addition, they explore data iteratively by testing various “what-if” scenarios. As a result, quick turnaround time for processing, searching, and exploring large document collections is critical. At the start of this project, no end-to-end scalable visual text analysis capabilities existed, and this prevented analysts from exploring, annotating, and analyzing large, existing document collections. The goal of this project was to couple Sandia’s world-class capabilities in high performance computing with expertise in text analysis and visualization.

We focused on the development of a suite of independent, scalable capabilities to process and search large document collections for use in data analysis and visualization software that could efficiently leverage parallel algorithms. Along the way, we developed exact and conceptual searching methods, as well as visual analysis methods, for understanding uncertainty inherent in models associated with text analysis. The resulting system, called ParaText, has served two purposes over the life of the project: (1) as an environment for rapid prototyping of algorithms , and (2) as a production capability for new and existing analysis software applications.

At the start of this project, much of the work at Sandia in text analysis had focused on entity and link extraction to populate graphs for visualization and graph-theoretic analysis. The ParaText LDRD project expanded on these ideas by using a combination of trees, vector spaces, and graphs to provide novel and distinguishing analysis and visualization capabilities to analysts interested in exploring textual data in a variety of ways not well suited to entity relationship graphs.

The primary data modeling technique researched and implemented in this project was Latent Semantic Analysis (LSA) [DeDuFuLaHa, DuFuLaDeHa, LaMcDeKi], where documents are modeled as term (feature) vectors. An LSA model is a statistical model of the variance of the terms both within and across the documents and can be used to describe the latent, or hidden, relationships between documents and terms appearing in those documents. LSA supplies conceptual organization and analysis of document collections by modeling high-dimension feature vectors in many fewer dimensions. In this project, we have concentrated on how to efficiently implement the LSA method and associated data processing when working with extremely large collections of documents.

Throughout this paper, we denote n as the number of documents in a collection, m as the number of unique terms, and p as the number of processors used for computation. LSA computes a truncated singular value decomposition (SVD) of a term-document matrix [BeDuOb], i.e., the collection of feature vectors associated with the documents in a text collection, or corpus. More specifically, the rank-*k* LSA model of a term-document matrix, $A \in \Re^{m \times n}$, is its rank-*k* SVD,

$$A_k = U_k \Sigma_k V_k^T, \tag{1}$$

where $U_k \in \Re^{m \times k}$, $\Sigma_k \in \Re^{k \times k}$, and $V_k \in \Re^{n \times k}$ contain the *k* leading left singular vectors, singular values, and right singular vectors, respectively. Furthermore, $U_k^T U_k = V_k^T V_k = I_k$, where I_k is the $k \times k$ identity matrix. Often, the rank of the LSA model in (1) is chosen such that *k* << min(*m,n*), leading to a reduction in model noise and computation for many analysis methods.

One particular type of analysis that is widely performed using LSA—and the motivating application for the work in the ParaText LDRD project—is determining conceptual relationships between two documents, two terms, or a term and a document. Graph data structures and algorithms are often used in this case [15]. For example, document clustering using graph layout methods and LSA modeling can be performed by first computing distances, or similarity scores, between all pairs of documents using the right singular vectors of the rank-*k* SVD of a term-document matrix. In this work, we use cosine similarities, defined as

$$e_{ij}(k) = \frac{\left\langle v_k^i \Sigma_k, v_k^j \Sigma_k \right\rangle}{\left\| v_k^i \Sigma_k \right\|_2 \left\| v_k^j \Sigma_k \right\|_2}, \tag{2}$$

between documents *i* and *j*, where $\langle \cdot,\cdot \rangle$ is the standard inner product, v_k^i is the *i*th row of V_k from (1), and $\| \cdot \|_2$ is the L^2-norm, or standard Euclidean norm. The similarities are stored as a similarity matrix, *E*, whose element (*i, j*) is defined in (2). To support large corpus analysis, only edge weights above a threshold are used in practice, leading to sparse similarity matrices. This similarity matrix is then used as a weighted adjacency matrix to construct a similarity graph. In this graph, nodes represent documents and edges represent the relationships between documents, weighted by similarity scores. Finally, graph layout methods are used to represent clusterings of the documents, i.e., related nodes are grouped together and unrelated nodes are separated in the resulting graph layout.

Much of our work focused on developing implementations of the ideas presented above for use on distributed memory systems. In the next section, we describe additional details of the full ParaText system. Finally, several additional research areas and associated software development efforts that were part of the ParaText LDRD project are described.

2.2. ParaText

The ParaText system is comprised of a collection of text analysis components designed to function within a Titan data processing pipeline [WyBa], where data sources, filters, and sinks can be combined in arbitrary ways. The ParaText components can be used as a C++, Python, or Java programming library, a set of command-line programs, or via a web service that implements a RESTful API [FiTa] atop a commodity HTTP server. Thus, the ParaText capabilities outlined in this report can be accessed using a variety of programming languages and environments.

2.2.1 The ParaText Pipeline

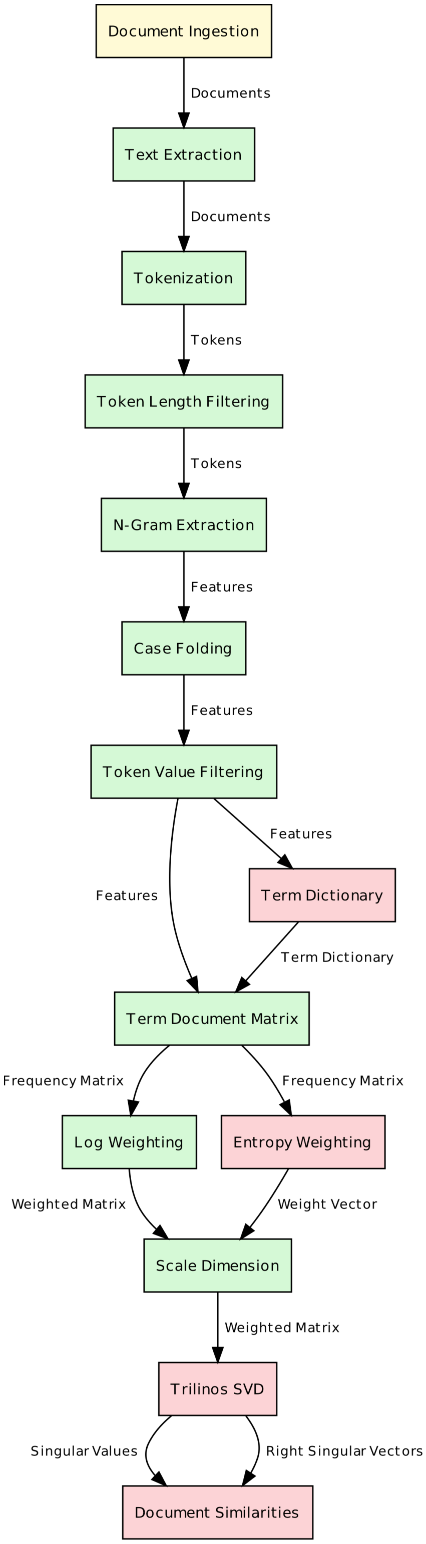


Figure 1: Typical ParaText Pipeline

The ParaText pipeline consists of a set of text-analysis and LSA-specific C++ components with inputs and outputs that are connected to form the pipeline. Because these components are automatically "wrapped" for use from other programming languages including Python and Java, users are free to create pipelines using their language of choice. In this section we outline a "typical" ParaText pipeline, while emphasizing that users may wish to alter the pipeline by substituting alternate components and/or configuring the pipeline differently.

The first part of the pipeline consists of filters for extracting and transforming text. With the exception of determining which files should be processed on which processors, the filters described in this section all parallelize extremely well.

Document Ingestion

The *Document Ingestion* filter is responsible for partitioning a set of documents and loading them into memory as a table where each row corresponds to a document. We have implemented several partitioning strategies that control how processors determine which files to load locally. The *Documents* partitioning strategy does a simple round-robin distribution where each process loads $1/p$ documents from the set. This strategy is simple to implement and requires no communication, but can lead to imbalanced loading as some processors may accumulate documents that are smaller- or larger-than-average. The *Bytes* partitioning strategy tries to balance loading by assigning files to processors so that each processor receives roughly the same number of input bytes. Because this is a variation on bin packing — a combinatorial NP-hard problem — we use a heuristic approach of maintaining a "bucket" for each processor, then inserting each file, in descending order of file size, into whichever bucket contains the fewest number of file bytes at the time. Early versions of this approach (which we call *Thrash*) did not require communication, but performed poorly due to file system contention as every processor simultaneously tried to retrieve the size of every file in the set. Subsequent versions use a single processor to retrieve file sizes and distribute them to the remaining processes before beginning the bucketing process.

Text Extraction

Once the local table of documents to be loaded has been created, we use MIME type information to extract text, using the *Text Extraction* filter. This filter contains a collection of strategy objects, each of which is responsible for extracting text from documents of a given MIME type. Note that the text extraction strategies can perform arbitrarily-complex operations to extract text from a document. These include extracting text from binary file formats such as PDF or word-processing documents, extracting metadata from images, or even performing optical character recognition on the contents of an image. For the experiments presented here, we relied on a default extraction strategy that handles all text/* MIME types. The extracted text is stored as Unicode [Un] strings using UTF-8 encoding, so the system is capable of working with mixed-language text.

Tokenization

Following text extraction, the *Tokenization* filter converts document text into a table of tokens. Tokenization is performed by splitting the document text into tokens using delimiters specified as half-open ranges of Unicode code points. Specifying ranges of logograms as "kept" delimiters supports tokenization and analysis of logosyllabic scripts such as Chinese, Korean, and Japanese, so that individual glyphs in those languages become tokens for analysis.

Token Length Filtering

We use two instances of the *Token Length* filter to discard tokens that are either too short or too long. This improves the downstream analysis by reducing noise in the data models.

N-Gram Extraction

The *N-Gram Extraction* filter converts individual tokens into n-grams and is parameterized to allow arbitrary values for n. We used unigrams (n = 1) for all experiments in this paper.

Case Folding

We use the *Case Folding* filter to transform the resulting tokens to a form where they can be used in case-insensitive comparisons. This transformation is carried out using the rules provided by Unicode, so the results can be used for case-insensitive comparisons across all Unicode-supported languages.

Token Value Filtering

To provide filtering of stop-words, we use the *Token Value* filter, which is parameterized by a list of tokens to be discarded. We used the standard stop word list from the SMART project [Vi].

Term Dictionary Creation

Once each processor has created its list of local terms (tokens), the *Term Dictionary* filter creates a global dictionary where each term is listed exactly once. Because this process necessitates communication of large numbers of strings between processors, we created several different implementations for testing: in *N-to-1*, every processor sends its local terms to processor 0, which creates the global dictionary and broadcasts the results back to every processor. For *N-to-N*, each processor broadcasts its local terms to all other processors, which then create their own copies of the global dictionary. In the *Binary Tree* approach, each processor sends its local terms to a "neighbor", which consolidates them with its own local terms, sending the results to a "super neighbor", and-so-on until the complete global dictionary has been created on one process that broadcasts the results to the others. The *Round Robin* approach involves processor k sending its local terms to processor $(k+1) \bmod p$, where they are consolidated with the local terms. This process runs p times, so that every term eventually reaches every processor. Finally, we implemented a *MapReduce* approach that uses the MapReduce-MPI library [PlDe] to consolidate and distribute terms.

Term Document Matrix Creation

Given the list of local terms and the global term dictionary computed by the *Term Dictionary* filter, each processor uses the *Term Document Matrix* filter to create its local portion of a sparse, distributed term-document frequency matrix (no inter-processor communication is required). For each term in the local term list, the global term dictionary is used to determine the corresponding matrix row. Two methods are implemented for term dictionary lookup: *Global* lookup is a naive approach where the global term dictionary is used to lookup each term with $O(m \log m)$ performance; *Global+Local* lookup is a more sophisticated two-stage approach where local lookup results are cached in a smaller lookup table for faster lookups.

Term Weighting

Once the term-document frequency matrix is generated, it must be weighted to incorporate the importance of the terms throughout the collection. In this paper, we focus on the standard log-entropy weighting scheme [EgLoBi] employed in many LSA studies, which illustrates the challenges associated with term weighting on distributed memory architectures. This weighting scheme involves the product of local quantities (frequencies of terms within each document) and global quantities (entropies of terms across the entire document collection). In ParaText, the local and global computations are separated into different filters: the *Log Weighting* and *Entropy Weighting* filters, respectively.

The entropy of term i across the collection is defined as

$$g_i = \frac{1}{n} \sum_{j=1}^{n} \frac{tf_{ij}}{gf_i} \log \frac{tf_{ij}}{gf_i}$$

where tf_{ij} is the frequency of term i in document j and gf_i is the global frequency of term i across the collection. Inter- processor communication is required both in computing gf_i for each term and the sum in g_i for each term. We have implemented several methods to study the impact of these communication requirements. In the *N-to-1* method, every processor computes its local values of gf_i and sends those to processor 0, which sums the values and broadcasts the results back to every processor. The sums for g_i are then computed in a similar fashion. In the *N-to-N* method, gf_i and g_i are first computed locally and then results are broadcast to all other processors for computing the global values. In both methods, there is the option to broadcast the locally computed values using either dense or sparse vectors. Once the local and global term weights are computed, the *Scale Dimension* filter then applies these weights to the matrix.

Singular Value Decomposition

To compute the SVD of the weighted term-document matrix, A, ParaText wraps the distributed block Krylov Schur method from the Anasazi package of the Trilinos solver library [BeHeLeTh]. Using shallow copies of data into the sparse matrix class in Trilinos, we avoid data replication. The rank-k truncated SVD of A is computed as $A_k = U_k \Sigma_k V_k^T$, where $U_k \in \Re^{m \times k}$, $\Sigma_k \in \Re^{k \times k}$, and $V_k \in \Re^{n \times k}$ are matrices containing the left singular vectors, singular values, and right singular vectors, respectively.

2.2.2 The ParaText Command Line Tools

While the ParaText pipeline presents a flexible system for research into LSA and related algorithms, we found that it provided more flexibility than was required for more production oriented environments where a user simply wishes to apply the standard LSA techniques to their data. Toward this end we have created two command-line tools, **paratext-lsa** and **paratext-lsa-query** that implement ParaText pipelines for generating LSA models and performing document search, respectively. Each tool can be run serially or in parallel using MPI, reading and writing model artifacts to the file system using a variety of file formats and parameters. For example, a user could compute an LSA model in serial on a collection of documents stored in a file system directory as follows:

```
$ paratext-lsa --directory /path/to/corpus –rank 25 --export-feature-
dictionary=features.vtk --export-global-weighting=weighting.vtk --export-
left-singular-vectors=lsv.vtk --export-singular-values=sv.vtk --export-right-
singular-vectors=rsv.vtk
```

This command creates a ParaText pipeline, ingests documents from the given file system directory, and computes a rank-25 LSA model, storing model artifacts to disk. Note that there are many more options for controlling the parameterization of the pipeline, exporting intermediate outputs, and importing pre-computed artifacts. For example, using the **paratext-lsa** executable a user could pre-compute a weighted term-document matrix once, storing it as an artifact, then use the stored matrix as an input to **paratext-lsa** in subsequent runs, computing multiple SVD models and bypassing the document parsing and tokenization stages of the pipeline.

Similarly, the artifacts generated by **paratext-lsa** are used as inputs to **paratext-lsa-query** when performing document search. For example, the following command uses the artifacts from the previous example to search for the terms “united” and “states” (and semantically similar terms):

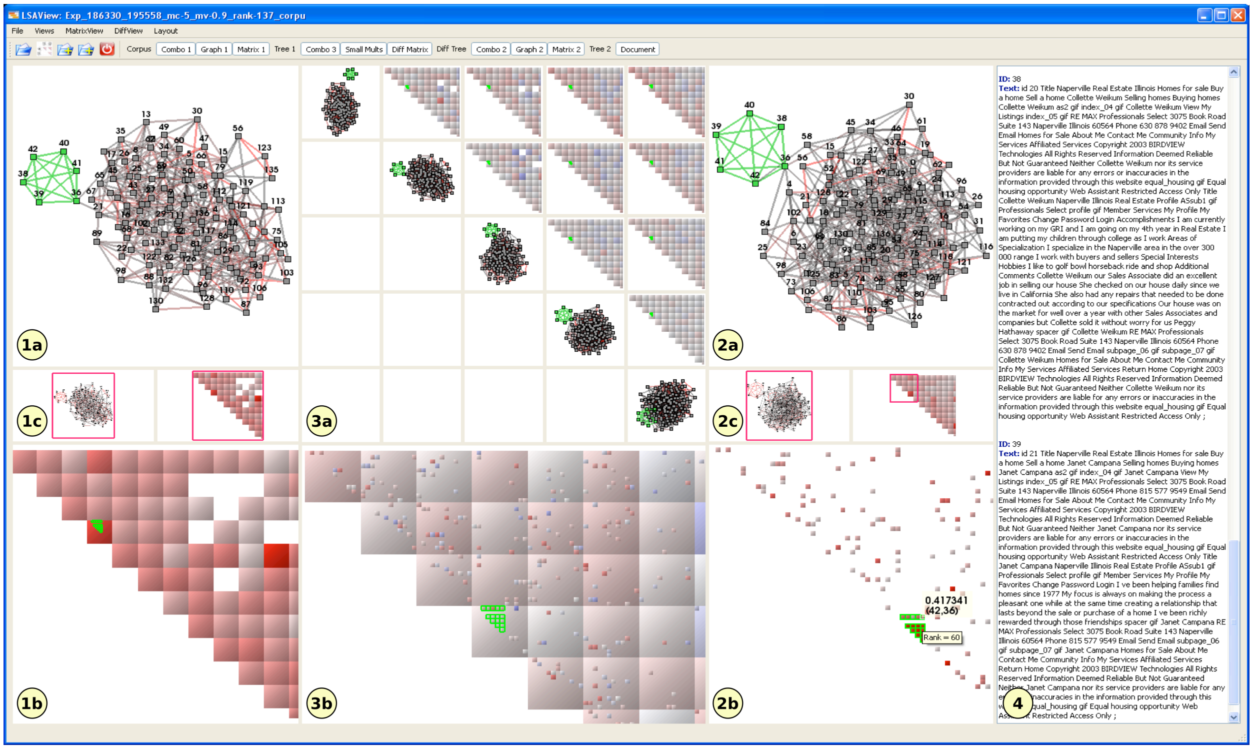
```
$ paratext-lsa-query --import-feature-dictionary=features.vtk --import-
global-weighting=weighting.vtk  --import-left-singular-vectors=lsv.vtk  --
import-singular-values=sv.vtk --import-right-singular-vectors=rsv.vtk --
query-text="united states" --export-similarity-table=similar-documents.vtk
```

2.2.3 ParaText Server

As a further means of using the ParaText system, we provide the ParaText server, which allows users to access the **paratext-lsa** and **paratext-lsa-query** executables through a RESTful HTTP interface via a commodity web server. This client-server approach to using ParaText makes it easier to integrate ParaText into production environments that already use commodity web protocols. Further, this client-server approach makes ParaText more accessible to end-users who are focused on analysis results rather than research. In its present form, ParaText Server is a shared-library module (plugin) for the popular Apache httpd server.

2.3. LSAView

LSAView was developed as a tool to interactively explore the impact of parameter choices on the model produced by LSA. We used the tool to examine different rank and scaling choices with respect to their impact on modeling and analysis functions from the perspective of the analyst at the end of the larger text analysis pipeline. Focusing on one of the central tasks that Latent Semantic Analysis is used for, calculating document similarities, we used document cluster structure to select the rank for a particular corpus. This value turned out to be much lower than the ranks recommended by conventional statistical approaches. A full account of this work is provided in our VAST paper [CrDuSh]. Figure 2 presents a screen shot of LSAView used to analyze two different LSA models of a single document collection to determine which model is more suitable for identifying document clusters.



ID: 38
Text: id 20 Title Naperville Real Estate Illinois Homes for sale Buy a home Sell a home Collette Weikum Selling homes Buying homes Collette Weikum as2 gif index_04 gif Collette Weikum View My Listings index_05 gif RE MAX Professionals Select 3075 Book Road Suite 143 Naperville Illinois 60564 Phone 630 878 9402 Email Send Email Homes for Sale About Me Contact Me Community Info My Services Affiliated Services Copyright 2003 BIRDVIEW Technologies All Rights Reserved Information Deemed Reliable But Not Guaranteed Neither Collette Weikum nor its service providers are liable for any errors or inaccuracies in the information provided through this website equal_housing gif Equal housing opportunity Web Assistant Restricted Access Only Title Collette Weikum Naperville Illinois Real Estate Profile ASsub1 gif Professionals Select profile gif Member Services My Profile My Favorites Change Password Login Accomplishments I am currently working on my GRI and I am going on my 4th year in Real Estate I am putting my children through college as I work Areas of Specialization I specialize in the Naperville area in the over 300 000 range I work with buyers and sellers Special Interests Hobbies I like to golf bowl horseback ride and shop Additional Comments Collette Weikum our Sales Associate did an excellent job in selling our house She checked on our house daily since we live in California She also had any repairs that needed to be done contracted out according to our specifications Our house was on the market for well over a year with other Sales Associates and companies but Collette sold it without worry for us Peggy Hathaway spacer gif Collette Weikum RE MAX Professionals Select 3075 Book Road Suite 143 Naperville Illinois 60564 Phone 630 878 9402 Email Send Email subpage_06 gif subpage_07 gif Collette Weikum Homes for Sale About Me Contact Me Community Info My Services Affiliated Services Return Home Copyright 2003 BIRDVIEW Technologies All Rights Reserved Information Deemed Reliable But Not Guaranteed Neither Collette Weikum nor its service providers are liable for any errors or inaccuracies in the information provided through this website equal_housing gif Equal housing opportunity Web Assistant Restricted Access Only ;

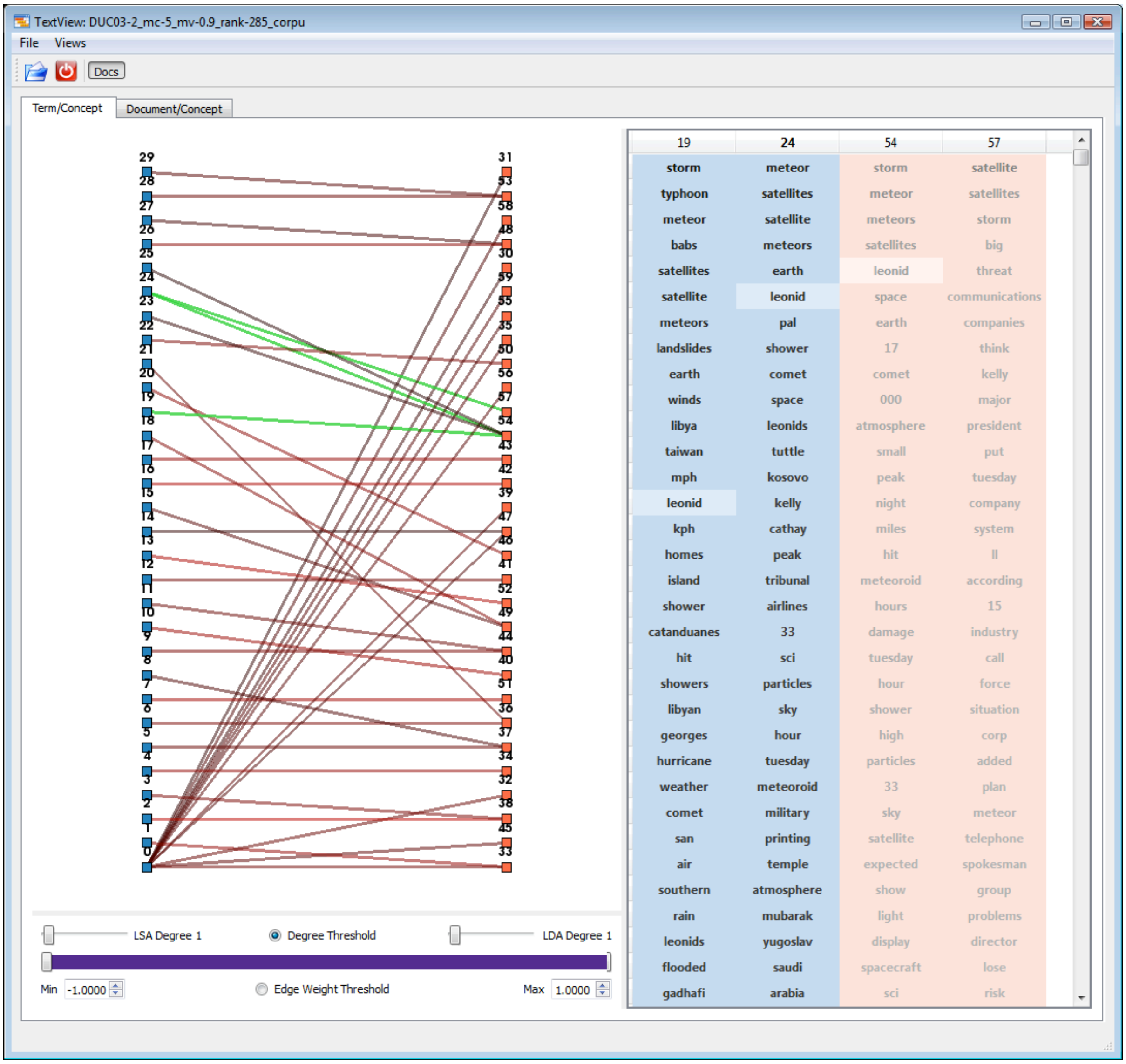
ID: 39
Text: id 21 Title Naperville Real Estate Illinois Homes for sale Buy a home Sell a home Janet Campana Selling homes Buying homes Janet Campana as2 gif index_04 gif Janet Campana View My Listings index_05 gif RE MAX Professionals Select 3075 Book Road Suite 143 Naperville Illinois 60564 Phone 815 577 9549 Email Send Email Homes for Sale About Me Contact Me Community Info My Services Affiliated Services Copyright 2003 BIRDVIEW Technologies All Rights Reserved Information Deemed Reliable But Not Guaranteed Neither Janet Campana nor its service providers are liable for any errors or inaccuracies in the information provided through this website equal_housing gif Equal housing opportunity Web Assistant Restricted Access Only Title Janet Campana Naperville Illinois Real Estate Profile ASsub1 gif Professionals Select profile gif Member Services My Profile My Favorites Change Password Login I ve been helping families find homes since 1977 My focus is always on making the process a pleasant one while at the same time creating a relationship that lasts beyond the sale or purchase of a home I ve been richly rewarded through those friendships spacer gif Janet Campana RE MAX Professionals Select 3075 Book Road Suite 143 Naperville Illinois 60564 Phone 815 577 9549 Email Send Email subpage_06 gif subpage_07 gif Janet Campana Homes for Sale About Me Contact Me Community Info My Services Affiliated Services Return Home Copyright 2003 BIRDVIEW Technologies All Rights Reserved Information Deemed Reliable But Not Guaranteed Neither Janet Campana nor its service providers are liable for any errors or inaccuracies in the information provided through this website equal_housing gif Equal housing opportunity Web Assistant Restricted Access Only ;

Figure 2. Examples of different views in the LSAView application: (1a) and (2a) Graph View, (1b) and (2b) Matrix View, (1c) and (2c) You Are Here View (3a) Small Multiples View, (3b) Difference Matrix View, and (4) Document View.

2.4. TextView

TextView is a visual tool for comparing LSA and Latent Dirichlet Analysis (LDA), a text analysis algorithm based on a probabilistic modeling approach. The algorithms are compared relative to the same input corpus with respect to two types of results, (1) the concepts identified and (2) the document similarity graphs produced. Each type of result is displayed in a separate tabbed view in the tool, as shown in the figures below. The user moves back and forth between views to explore the connections between topics, terms, and documents.

Under the *Term/Concept* tab (Figure 3), a bipartite graph displays the similarities between LSA concepts (blue nodes on the left) and LDA topics (orange nodes on the right). The strength of the similarity between any concept-topic pair is color-coded in the edge between them, with the color range going from blue to black to red. To the right of the graph is a table with the terms for each concept (light-blue columns) or topic (light-orange columns) in sorted order of importance. The strength of each term's contribution to a particular concept/topic is also depicted by the darkness of the font used to display it. Selecting edges in the bipartite graph will down-select the table to just the concept/topic pairs connected by those edges. A term within the table of concept/topic lists can be selected, highlighting that same term in all lists and within any displayed document texts in the *Document/Concept* tab (Figure 4).



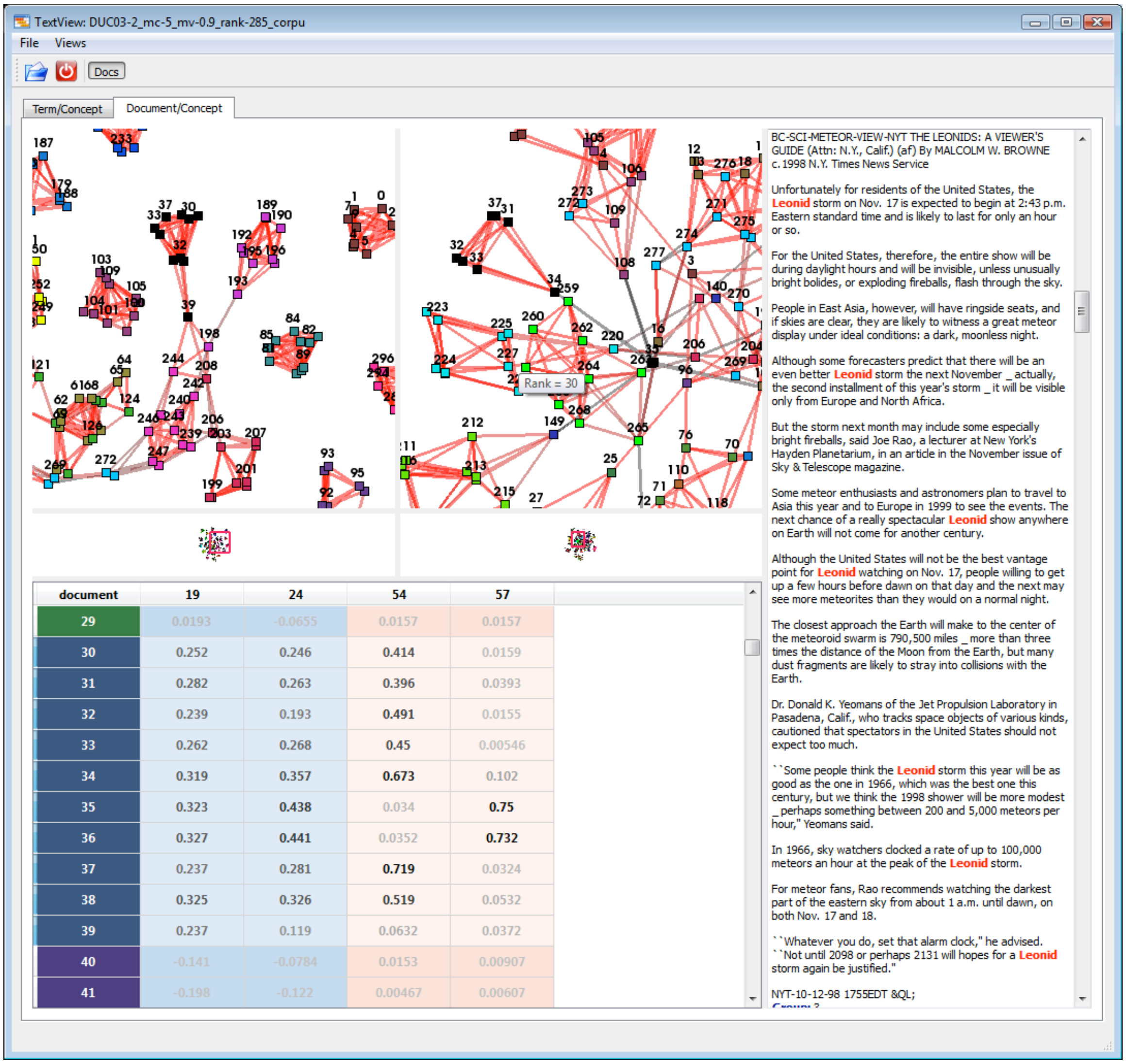
19	24	54	57
storm	meteor	storm	satellite
typhoon	satellites	meteor	satellites
meteor	satellite	meteors	storm
babs	meteors	satellites	big
satellites	earth	leonid	threat
satellite	leonid	space	communications
meteors	pal	earth	companies
landslides	shower	17	think
earth	comet	comet	kelly
winds	space	000	major
libya	leonids	atmosphere	president
taiwan	tuttle	small	put
mph	kosovo	peak	tuesday
leonid	kelly	night	company
kph	cathay	miles	system
homes	peak	hit	ll
island	tribunal	meteoroid	according
shower	airlines	hours	15
catanduanes	33	damage	industry
hit	sci	tuesday	call
showers	particles	hour	force
libyan	sky	shower	situation
georges	hour	high	corp
hurricane	tuesday	particles	added
weather	meteoroid	33	plan
comet	military	sky	meteor
san	printing	satellite	telephone
air	temple	expected	spokesman
southern	atmosphere	show	group
rain	mubarak	light	problems
leonids	yugoslav	display	director
flooded	saudi	spacecraft	lose
gadhafi	arabia	sci	risk

LSA Degree 1 ● Degree Threshold LDA Degree 1

Min -1.0000 ○ Edge Weight Threshold Max 1.0000

Figure 3: Term/Concept tab views with LSA concepts in light-blue and LDA topics in light-orange. The bipartite graph displays weighted similarities between LSA concepts on the left and LDA topics on the right. Importance-ordered term lists for concepts and topics are in the term/topic table on the right.

The *Document/Concept* tab views include document similarity graphs for LSA (left) and LDA (right), in which the nodes represent documents and the edges display the strength of the similarity between them. Nodes are color-coded according to human-defined concept labels provided by the data source, a summarization contest. Below each graph is a *You Are Here* view, which shows the view outline as a red rectangle within the context of an overview of the full graph (see the center views in Figure 4). Below that is a table displaying the weighted contributions for each document for each concept or topic. The background color for the document identifier uses the same label color-coding as the nodes to assist in finding the associated nodes in the graphs. Selecting documents from either the table or the graphs will display the documents' text in the view on the right.



document	19	24	54	57
29	0.0193	-0.0655	0.0157	0.0157
30	0.252	0.246	0.414	0.0159
31	0.282	0.263	0.396	0.0393
32	0.239	0.193	0.491	0.0155
33	0.262	0.268	0.45	0.00546
34	0.319	0.357	0.673	0.102
35	0.323	0.438	0.034	0.75
36	0.327	0.441	0.0352	0.732
37	0.237	0.281	0.719	0.0324
38	0.325	0.326	0.519	0.0532
39	0.237	0.119	0.0632	0.0372
40	-0.141	-0.0784	0.0153	0.00907
41	-0.198	-0.122	0.00467	0.00607

BC-SCI-METEOR-VIEW-NYT THE LEONIDS: A VIEWER'S GUIDE (Attn: N.Y., Calif.) (af) By MALCOLM W. BROWNE
c. 1998 N.Y. Times News Service

Unfortunately for residents of the United States, the Leonid storm on Nov. 17 is expected to begin at 2:43 p.m. Eastern standard time and is likely to last for only an hour or so.

For the United States, therefore, the entire show will be during daylight hours and will be invisible, unless unusually bright bolides, or exploding fireballs, flash through the sky.

People in East Asia, however, will have ringside seats, and if skies are clear, they are likely to witness a great meteor display under ideal conditions: a dark, moonless night.

Although some forecasters predict that there will be an even better Leonid storm the next November _ actually, the second installment of this year's storm _ it will be visible only from Europe and North Africa.

But the storm next month may include some especially bright fireballs, said Joe Rao, a lecturer at New York's Hayden Planetarium, in an article in the November issue of Sky & Telescope magazine.

Some meteor enthusiasts and astronomers plan to travel to Asia this year and to Europe in 1999 to see the events. The next chance of a really spectacular Leonid show anywhere on Earth will not come for another century.

Although the United States will not be the best vantage point for Leonid watching on Nov. 17, people willing to get up a few hours before dawn on that day and the next may see more meteorites than they would on a normal night.

The closest approach the Earth will make to the center of the meteoroid swarm is 790,500 miles _ more than three times the distance of the Moon from the Earth, but many dust fragments are likely to stray into collisions with the Earth.

Dr. Donald K. Yeomans of the Jet Propulsion Laboratory in Pasadena, Calif., who tracks space objects of various kinds, cautioned that spectators in the United States should not expect too much.

``Some people think the Leonid storm this year will be as good as the one in 1966, which was the best one this century, but we think the 1998 shower will be more modest _ perhaps something between 200 and 5,000 meteors per hour," Yeomans said.

In 1966, sky watchers clocked a rate of up to 100,000 meteors an hour at the peak of the Leonid storm.

For meteor fans, Rao recommends watching the darkest part of the eastern sky from about 1 a.m. until dawn, on both Nov. 17 and 18.

``Whatever you do, set that alarm clock," he advised.
``Not until 2098 or perhaps 2131 will hopes for a Leonid storm again be justified."

NYT-10-12-98 1755EDT &QL;

Figure 4: Document/Concept tab views: document similarity graphs (LSA on the left, LDA on the right), You Are Here views below each graph providing context, a table of document weights for each concept/topic, and document text highlighting the selected term in the term/topic table. Selected documents are shown in black in the graphs, highlighted in the table, and their text displayed.

2.5. HEMLOCK

HEMLOCK is a software tool for constructing, evaluating, and applying heterogeneous ensemble data models for use in solving classification problems involving data with continuous or discrete features. HEMLOCK consists of various data readers, machine learning algorithms, model combination and comparison routines, evaluation methods for model performance testing, and interfaces to external, state-of-the-art machine learning software libraries. HEMLOCK uses XML for all input and output, and standard readers and writers are being used for data input and output. Data models are created by a variety of supervised learning methods: decision tree and random forest inducers plus a linear perceptron learner as part of HEMLOCK along with interfaces to the methods available in the WEKA software library of machine learning algorithms. Evaluation methods for assessing individual model performance include accuracy computation, confusion matrix generation, receiver operating characteristics (ROC) analysis, and area under the curve (AUC) analysis. Methods for combining heterogeneous models into a single ensemble model include majority voting and parameter regression.

Classification is the task of learning a target function that maps data instances to one of several predefined categories. These target functions are also called classifiers, classifier models, and hypotheses. We refer to a classifier constructed or learned from an ensemble of different types of classifiers is as a heterogeneous ensemble classifier. Note that such classifier models are also referred to as hybrid ensemble classifiers. There are several challenges associated with learning heterogeneous ensemble classifiers. The choice of base classifiers (i.e., ensemble member classifier models) needs to be determined. Classifier performance can differ greatly across data sets, and thus choosing the collection of classifiers that will best classify a given set of data is often a difficult task. Each base classifier can be parameterized in many different ways, and thus an understanding of how these parameters are correlated within each base classifier as well as across the ensemble is key to classifying data sets accurately.

A further challenge is combining base classifiers effectively, so that the performance of the ensemble classifier is better than that of the individual classifiers. There are two basic strategies for combining classifiers in an ensemble: fusion and selection [WoBoKe]. Ensembles that use selection try to find the best classifier ensemble member that is most capable of correctly classifying a particular instance. Ensembles that use selection are also known as cooperative ensembles. In contrast to selection, fusion methods make use of the outputs of all of the classifiers to try to determine the label of an instance. Voting is an example of fusion: each of the classifiers in the ensemble is given one vote and all of the votes are counted towards deciding which output label should be chosen. Ensembles that use fusion are commonly referred to as competitive ensembles. There are three levels at which classifiers output can be combined using fusion: label, ranking, measurement. At the label level the ensemble will only use the one class label that each of the base classifiers determines is correct. For ranking, base classifiers in the ensemble provide a ranked list of class labels reflecting how likely each class is marked as the correct label for each data instance. Finally, at the measurement level each of the base classifiers provides output that is intrinsic to the particular learning algorithm used. Typically, measurements consist of probability distributions of the class assignment for each instance.

It has been shown that the strength of an ensemble is related to the performance of the base classifiers and the lack of correlation between them (i.e., model diversity) [BiWa, WaPaEt]. One way to decrease the correlations between the classifiers while increasing or maintaining the overall performance of the ensemble classifier is to include base classifiers derived from different learning algorithms such as decision trees, neural networks, perceptrons, support vector machine, etc.

Figure 5 illustrates the effectiveness of using HEMLOCK to build improved heterogeneous ensemble models. The task being evaluated was handwritten digit classification, i.e., trying to determine which digit is represented in a scanned image of a digit written by a human. The figure presents results of using three different classification modeling methods: association rules, a naïve Bayes classifier, and a decision tree. Moreover, the results include both homogeneous ensemble classifiers and the best performing heterogeneous ensemble classifier (chosen from a collection of ensemble classifiers created using various combinations of the individual classifier models). The results indicate that the heterogeneous ensemble models outperform both the individual classifiers as well as the homogeneous ensemble classifiers. Although there was more computation involved in the heterogeneous ensemble classifiers (due to the many models built in fusing the individual classifiers in different ways), the improved models required no subject matter expertise and were generated automatically. HEMLOCK thus provides new approaches to problems and research involving supervised machine learning and empirical predictive modeling.

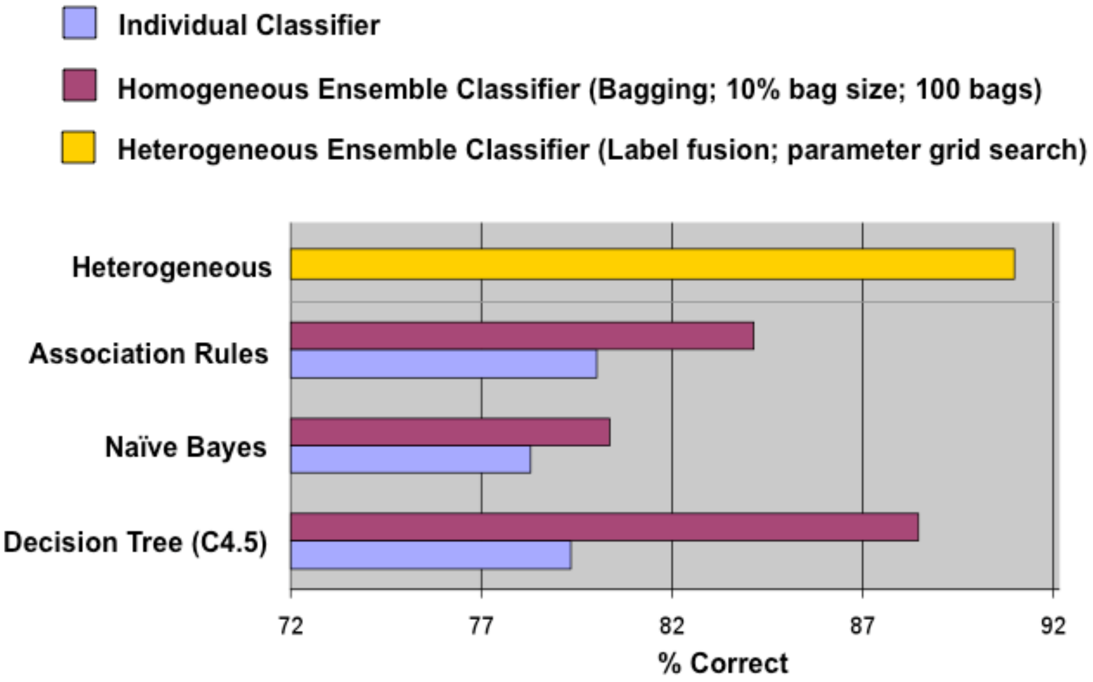


Figure 5. Performance comparisons between base classifiers, homogeneous ensembles, and heterogeneous ensembles for the task of image classification.

3. IMPACT

In this section, we describe the project's technical, application, and the programmatic impacts.

3.1 Technical Impact

In addition to the direct impact of the ParaText research itself, the ParaText project has had tremendous influence on projects both internal and external to Sandia National Laboratories.

- As part of the initial release of the open source version of the Titan toolkit, the ParaText filters, utilities, and support components for scalable text analysis became the first fully operational Titan Library.

- To make it easier to use the Trilinos linear algebra libraries (also developed at Sandia) to implement the SVD calculations in ParaText, the ParaText team converted a subset of the Trilinos libraries to the CMake [MaHo] build system. Based on the success of this effort, the Trilinos team later converted the entire Trilinos project to CMake. This accomplishment paved the way for Titan/VTK users and developers to use the powerful distributed memory linear and nonlinear solvers in Trilinos and provided native configuration and build capabilities on desktop platforms (Windows and MacOS) not previously supported by Trilinos.

- To support the linear algebra operations needed for the ParaText pipeline, the ParaText team developed sparse and dense arbitrary-dimension array data structures for Sandia's Titan framework. These data structures have become a key component in Titan, and have been used by other projects including the Networks Grand Challenge LDRD.

- To support ParaText Server, the ParaText team developed expertise and experience around commodity internet protocols (HTTP, RESTful APIs, etc) that is having a profound impact on Sandia data analysis across domains from intelligence analysis to scientific visualization.

- To expand the set of features computed in the ParaText models beyond space-delimited terms, the ParaText team developed an interface to the Linguistica software library[1] for morphological analysis. The benefits of using morphemes (i.e.,) over space-delimited terms in information retrieval applications by members of the ParaText team [ChBaAb] led to integration of Linguistica routines for morphological analysis into Titan.

3.2 Project and Application Impact

3.2.1 Networks Grand Challenge (NGC)

Many of the techniques and components developed as part of the ParaText project have been used by the Networks Grand Challenge (NGC) LDRD project. Examples include the sparse and

[1] http://linguistica.uchicago.edu/linguistica.html

dense array data structures developed for ParaText (repurposed for tensor analysis by the NGC), the ParaText LSA pipeline (combined with graph analysis techniques for one of the NGC prototypes) and the ParaText approach to client-server functionality (expanded by the NGC into the realm of interactive web visualization).

3.2.2 ThreatView

To support national security applications involving document analysis, including conceptual information retrieval (as opposed to Boolean search), cluster analysis, and entity relationship analysis, the ParaText team developed the Algebraic Engine. The Algebraic Engine is an interface to the LSALIB text analysis library and serial implementations of the ParaText pipeline components that supports caching for efficient use of multiple document collections, text models, and analysis artifacts. The main use of the Algebraic Engine is in the ThreatView information visualization tool.

3.2.3 Nuclear Attribution

The Timeline Treemap Browser was an application developed by Sandia National Laboratories for the National Technical Nuclear Forensics Center (NTNFC), in the Domestic Nuclear Detection Office (DNDO). The tool was part of a larger project in nuclear materials forensics for tracing material samples back to production sources. The browser enabled timelines of nuclear material production information from various facilities to be combined and explored. The application used LSALib to analyze documents associated with the timelines, producing an alternate trend-based view of the material production information that was visualized as a stacked chart showing how the relative importance of various themes in the documents changed over time. The application is described in a SAND report [CrHu], though the report is OUO.

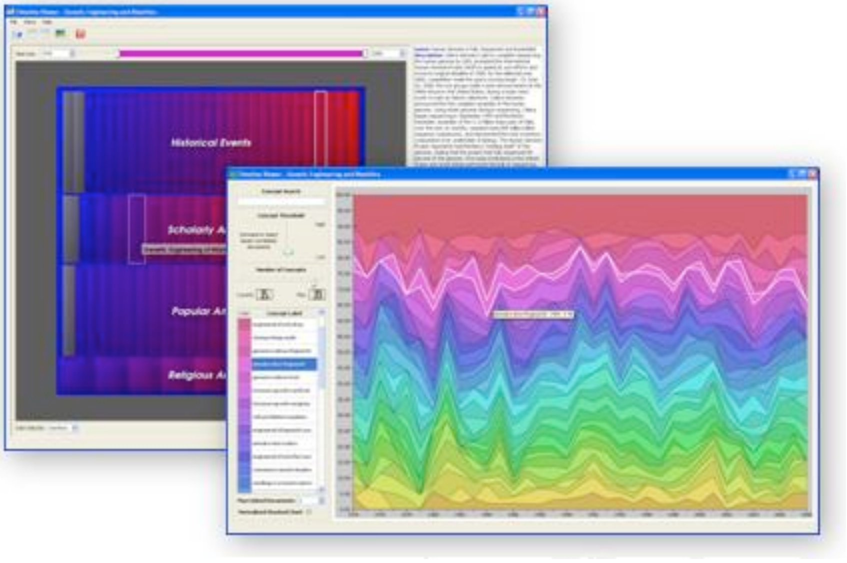
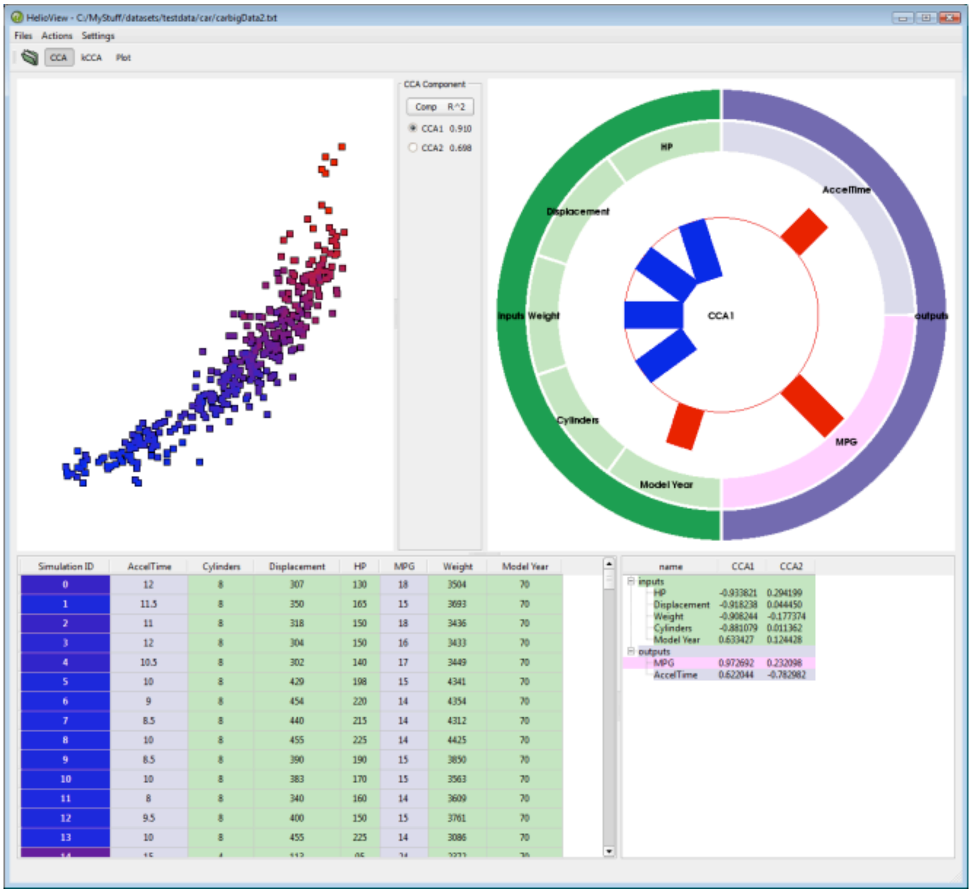


Figure 6: Views of the Timeline Treemap Browser displaying the evolution of genetic engineering data set. The back image shows the treemap view of timeline elements. The front image is the trend-based view showing thematic changes over time.

3.2.4 ParaSpace

ParaSpace is a CSRF-funded project to perform scalable sensitivity analysis on ensembles of simulation runs. It began as a generalization of ParaText's vector space model, originally intending to replace the bag-of-words used in Latent Semantic Analysis (LSA) with a bag-of-features taken from a combination of input parameters and output features extracted from finite element mesh files generated by the simulations. In this way, the LSA term/document matrix was to be replaced by a feature/simulation matrix that could be passed directly into ParaText's scalable infrastructure for doing Singular Value Decomposition (SVD). However, as the project progressed Canonical Component Analysis (CCA) was found to be better suited to correlating inputs and outputs. Although this means that ParaSpace can not use the ParaText pipeline directly, the ParaSpace CCA can still benefit from the ParaText SVD implementation, and is based upon ParaText components. Next year ParaSpace will use even more of the ParaText code-base, building upon the ParaText Server to convert ParaSpace to a multi-tier web-based delivery model. Shown in Figure 7, HelioView is a prototype application that was developed as part of the project for exploring correlations between input parameters and output metrics (non-mesh results).



Comp	R^2
● CCA1	0.910
○ CCA2	0.698

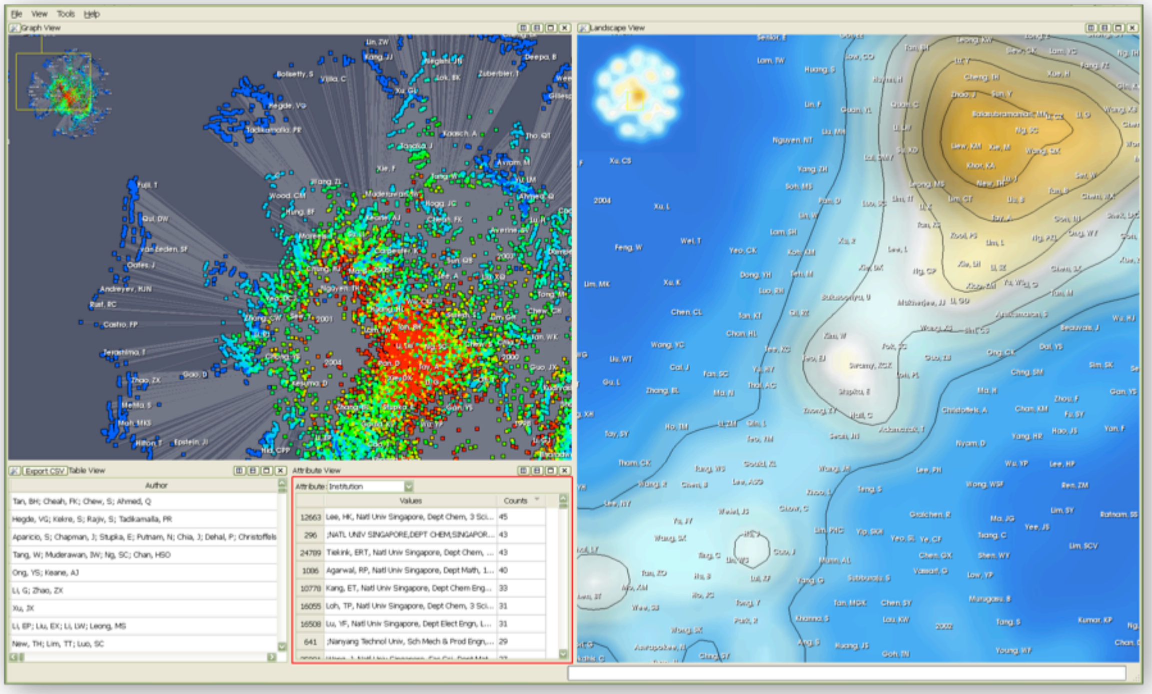
Simulation ID	AccelTime	Cylinders	Displacement	HP	MPG	Weight	Model Year
0	12	8	307	130	18	3504	70
1	11.5	8	350	165	15	3693	70
2	11	8	318	150	18	3436	70
3	12	8	304	150	16	3433	70
4	10.5	8	302	140	17	3449	70
5	10	8	429	198	15	4341	70
6	9	8	454	220	14	4354	70
7	8.5	8	440	215	14	4312	70
8	10	8	455	225	14	4425	70
9	8.5	8	390	190	15	3850	70
10	10	8	383	170	15	3563	70
11	8	8	340	160	14	3609	70
12	9.5	8	400	150	15	3761	70
13	10	8	455	225	14	3086	70

name	CCA1	CCA2
⊟ inputs		
HP	-0.933821	0.294199
Displacement	-0.918238	0.044450
Weight	-0.908244	-0.177374
Cylinders	-0.881079	0.011362
Model Year	0.633427	0.124428
⊟ outputs		
MPG	0.972692	0.232098
AccelTime	0.622044	-0.782982

Figure 7: HelioView application showing CCA correlations for car data set.

3.2.5 LDRDView

A software application written for the Sandia LDRD office to support analysis of their funding portfolio, LDRDView was developed before this project started. One of the initial accomplishments of this project was interfacing LDRDView with the LSALIB text analysis library. This increased the amount of data that could analyzed using LDRDView and provided new capabilities for conceptual search and relationship analysis between projects in the LDRD portfolio. Figure 8 presents a snapshot of the LDRDView application using the tools developed as part of this project.



Author
Tan, BH; Cheah, FK; Chew, S; Ahmed, Q
Hegde, VG; Kekre, S; Rajiv, S; Tadikamalla, PR
Aparicio, S; Chapman, J; Stupka, E; Putnam, N; Chia, J; Dehal, P; Christoffels
Tang, W; Muderawan, IW; Ng, SC; Chan, HSO
Ong, YS; Keane, AJ
Li, G; Zhao, ZX
Xu, JX
Li, EP; Liu, EX; Li, LW; Leong, MS
New, TH; Lim, TT; Luo, SC

Attribute: Institution

	Values	Counts
12663	Lee, HK, Natl Univ Singapore, Dept Chem, 3 Sci...	45
296	;NATL UNIV SINGAPORE,DEPT CHEM,SINGAPOR...	43
24709	Tiekink, ERT, Natl Univ Singapore, Dept Chem, ...	43
1006	Agarwal, RP, Natl Univ Singapore, Dept Math, 1...	40
10778	Kang, ET, Natl Univ Singapore, Dept Chem Eng...	33
16055	Loh, TP, Natl Univ Singapore, Dept Chem, 3 Sci...	31
16508	Lu, YF, Natl Univ Singapore, Dept Elect Engn, L...	31
641	;Nanyang Technol Univ, Sch Mech & Prod Engn,...	29

Figure 8. The LDRDView application for analyzing funding portfolios.

3.2.6 Multi-role Experiential Learning

ParaText was used in the experimental studies that were part of the Real-Time Individualized Training Vectors for Experiential Learning LDRD project. Specifically, benefits of using LSA to model participants for analyzing interactions in multi-role online environment were demonstrated in a recent peer-reviewed conference publication [RaFaTuWi].

3.3 Programmatic Impact

Programmatic impact of the ParaText LDRD project includes extensions of ideas leading to new funding proposals, activities of leadership and service by ParaText team members as a direct result of involvement on the project, several workshops associated with text analysis and visualization led by ParaText team members, several awards, and support for several summer intern students.

3.3.1 Funding

The following sections describe the impact of the ParaText LDRD project in terms of proposals submitted to a wide variety of sources. Analysis ideas and software capabilities developed as part of the ParaText LDRD project directly impacted these proposals.

3.3.1.1 CSRF/CSSE

The following CSRF/CSSE project was a direct consequence of the work on the ParaText LDRD project extended to analyzing modeling and simulation data associated with the ASC Program.

- ***Sensitivity Analysis and Relationship Discovery in Large Ensembles of Simulation Runs (ParaSpace)***
  Patricia Crossno (PI), Timothy Shead (team), Shawn Martin, Warren Hunt, FY09-FY11, funded, $1.4M. This project extends the vector space model of ParaText from text corpora to finding correlations between inputs and outputs in ensembles of finite element simulations runs for doing sensitivity analysis.

3.3.1.2 DOE Office of Science

The following proposals submitted to the DOE Office of Science feature ideas originating from the ParaText project, including large-scale text analysis and machine learning.

- ***Mathematical Analysis of Imperfect Petascale Data***
  Daniel Dunlavy (PI), FY10-FY12, ASCR MAPD call, not funded
  We propose to develop novel mathematical techniques for addressing data imperfections in the construction of feature data models and analysis methods. Our main goal is to develop methods for building data compression and machine learning models where local models are moved to distributed data stores and then updated, thus avoiding any moving of raw data. Local models will approximate global characteristics of the data as the models are passed and updated across the data stores, and techniques from ensemble learning will be used to optimally utilize local models to facilitate global analysis. Data imperfections will be incorporated into the modeling process directly and propagated across the data stores along with the analysis models to enable understanding of their impact on the overall characteristics of the data.

- ***Topological Sensitivity Analysis for Text Analysis over Complex Networks***
  Scott Mitchell (PI), Daniel Dunlavy (team), FY10-FY12, ASCR MAPD call, not funded
  We propose to study combinatorial algebraic topology, specifically simplicial homology groups and their generators, Morse theory and Reeb graphs, and their effective computation. These components are central to understanding the space of possible interpretations of a corpus of text documents distributed over a computer network.

- ***Institute for Informatics, Discrete Systems, and Data Analytics (IIDSDA)***
  Richard Murphy (PI), Daniel Dunlavy (team), FY10-FY12, ASCR Math/CS Inst., not funded
  To address the challenges of enabling peta- to extreme-scale computations for these new classes of problems, we propose to establish the DOE Institute for Informatics, Discrete Systems, and Data Analytics (IIDSDA). The institute will have a unified, multi-institutional structure consisting of members in the national labs, universities, and industry, with interaction and outreach across all three sectors.

3.3.1.3 LDRD

The following LDRD proposals contain data analysis and software ideas that stemmed from the ParaText LDRD project. Information listed for each proposal includes, project title, ParaText project staff associated with the proposed work, dates, LDRD IAT where the proposal was submitted, funding status, and a short description of the proposed work. Although none of the proposals were funded, the depth and variety of ideas as well as the IATs to which the work was proposed (EPS, DSA, NW and Senior's Council) indicates the broad impact of the ParaText project on future work in the area of text analysis.

- ***Uncertainty Analysis of Large Complex Text Modeling and Analysis Data Flow Pipelines***
  Daniel Dunlavy (PI), Timothy Shead (team), FY11-FY13, EPS, full proposal, not funded
  We propose development of new methodologies based on probabilistic and topological analysis to quantify the propagation of uncertainties through complex text modeling and analysis data flow pipelines; application of these methodologies to problems in the areas of cybersecurity and nuclear nonproliferation.

- ***Assessing Terrorist Cell Viability from Internet Discourse***
  Link Hamilton (PI), Daniel Dunlavy (team), FY11-FY13, DSA, idea, not funded
  We propose to develop text mining techniques in order to identify recruitment messages, leaders, and followers from on-line conversations between gamers, and to follow the evolution of the groups to study dynamic properties like team lifetimes and organization life cycles.

- ***Machine Learning for Adaptive Vulnerability Detection in Software Source Code***
  Justin Basilico (PI), Daniel Dunlavy (team), FY11-FY13, DSA, idea, not funded
  We propose applying supervised and unsupervised techniques of machine learning for structured data, such as statistical relational learning, hidden Markov models, and kernel-based methods, to the variety of structural relationships exhibited in programs and in combination with textual features.

- ***Advanced Text Processing for Training Applications***
  Andrew Scholand (PI), Daniel Dunlavy/Timothy Shead (team), FY11-12, EPS, not funded
  We propose to conduct an in-depth analysis of the educational training data with linguistic and other statistical tools not currently available in Titan, evaluate the utility of the resulting performance assessment, and where appropriate and valuable, incorporate these educational data mining tools into Titan.

- ***An Integrated Approach to e-Discovery, Email Marking, and Records Retention: Improving Workforce Compliance and Efficiency Using Text Analysis***
  J.T. McClain (PI), Daniel Dunlavy (team), FY10-FY12, Seniors' Council, idea, not funded
  We will build upon toolsets developed by the Cognitive Systems departments to develop a system to categorize emails on the user's desktop, provide visual cues with suggestions on marking and retention, and aide in following through with the policy requirements (marking, retention, distribution, etc).

- ***Integration, Analysis and Visualization of Data and Information Relationships for Strengthening our Knowledge and Confidence in the U.S. Stockpile***
  Daniel Dunlavy (PI), Patricia Crossno (team), FY09-FY11, NW, idea, not funded
  We will leverage the Sandia-developed TITAN information visualization toolkit to develop new data analysis and visualization methods to solve the problem of integration and analysis of data associated with stockpile stewardship and planning.

- ***Anticipating Technology Innovation***
  Patricia Crossno (PI), Daniel Dunlavy (team), FY09-FY11, DSA, idea, not funded
  The proposal combined text analysis of technology trends in the research literature with a learning classifier system to categorize concepts according to the Theory of Inventive Problem Solving (TRIZ).

3.3.1.4 NMSBA

The New Mexico Small Business Assistance (NMSBA) program allows Sandia National Laboratories and Los Alamos National Laboratory to use a portion of their gross receipts taxes paid each year to provide technical advice and assistance to New Mexico small businesses. Through this program, lab researchers can help small businesses address important challenges to their business by using laboratory resources at no cost to the small business. The projects listed below were a direct consequence of the work on the HEMLOCK software framework developed as part of the ParaText LDRD project.

- ***Improved Outlier/Anomaly Detection***
  Daniel Dunlavy (PI), FY10, funded, $10K
  Development of methods for model-based outliers associated with diabetes screening applications.

- ***Dynamic Model Selection for Regression Models***
  Daniel Dunlavy (PI), FY11, in review as of September 2010
  Development of patient-specific selection methods for regression-based prediction models associated with diabetes screening applications.

3.3.2 Leadership and Service

Table 1 lists the leadership and service responsibilities associated with the ParaText LDRD project. All of the items listed came as a direct result of the work on this project.

Table 1. Leadership and service associated with the ParaText LDRD project.

Date(s)	Description	Personnel
6/08–present	Trilinos Advisory Group Member (direct consequence of ParaText-Trilinos interactions)	Daniel Dunlavy
7/08	Workshop Organizer: 2008 Sandia Workshop on Data Mining and Data Analysis	Daniel Dunlavy
9/09	Reviewer, IEEE Transactions on Visualization and Computer Graphics	Patricia Crossno
8/09	Reviewer, 5th International Symposium on Visual Computing	Tim Shead
3/08–7/08	Program Committee and Reviewer, IEEE Visualization 2008	Patricia Crossno
3/09–7/09	Program Committee and Reviewer, IEEE Visualization 2009	Patricia Crossno
3/10–7/10	Program Committee and Reviewer, IEEE Visualization 2010	Patricia Crossno
FY09	Analytics Advisory Team (SNL search capabilities leveraging informatics research)	Daniel Dunlavy
10/08–3/09	Organizer, MAPD (Mathematics for Petascale Data Analysis) Working Group	Daniel Dunlavy
2/09	Review Panel Member*, National Science Foundation, Department of Mathematics*	Daniel Dunlavy
2/08–7/10	SIAM Professional Development Committee	Daniel Dunlavy
3/09	Co-organizer (with Misha Kilmer of Tufts University) for the Professional Development Evening of the 2009 SIAM Conference on Computational Science and Engineering Meeting	Daniel Dunlavy
10/08–5/09	Co-organizers, VizMining 2009 Workshop, accepted at the 2009 SIAM International Conference on Data Mining, Sparks, NV	Daniel Dunlavy Patricia Crossno Tim Shead
3/09–7/09	Co-organizer (with Suzanne Shontz of Penn State University), Professional Development Evening, 2009 SIAM Annual Meeting	Daniel Dunlavy
3/10-7/10	Co-organizer (with Suzanne Shontz of Penn State University), Professional Development Evening, 2010 SIAM Annual Meeting	Daniel Dunlavy
4/10-6/10	Program Committee and Reviewer, 2nd Workshop on Large-scale Data Mining: Theory and Applications	Daniel Dunlavy
5/10	Reviewer, IEEE Visualization 2010	Tim Shead
8/09	Reviewer, International Symposium on Visual Computing 2009	Tim Shead
8/10	Reviewer, International Symposium on Visual Computing 2010	Tim Shead

3.3.3 Workshops

As part of the ParaText LDRD project, two workshops primarily focused on data analysis (and significantly on text analysis) and visualization were organized. The first, the 2008 Sandia Workshop on Data Mining and Data Analysis, was a follow-on to the 2007 workshop of the same name. It bought together researchers and application users interested in an open dialogue about data analysis problems and solutions in the current environment at Sandia National Laboratories. The second, the 2009 VisMining Workshop, was organized as part of the SIAM International Conference on Data Mining (SDM). Although the workshop was accepted by the SDM Program Committee and had several prominent researchers in the areas of data analysis and visualization on the VizMining Program Committee, a lack of quality submitted papers led to a cancellation of the workshop.

3.3.3.1 WDMDA

The 2008 Sandia Workshop on Data Mining and Data Analysis was co-organized by Daniel Dunlavy, Jim Brandt, and Ann Gentile. The intent of the workshop is to discuss research ideas, technical challenges, open questions, and potential for collaboration across departments, centers, and applications at Sandia in the areas of data mining and data analysis listed below. The workshop will comprise presentations by researchers from Sandia in these areas and group discussions, with the intent of understanding what we are doing today, how it fits into the world at large, and what seems promising to tackle next. The one-day workshop included over 50 participants and 12 presentations. A SAND report including a summary of the workshop findings and extended abstracts of the talks is available [BrDuGe].

3.3.3.2 VizMining

The VizMining 2009 Workshop was to be held in conjunction with the 2009 SIAM International Conference on Data Mining. The goal of this workshop was to bring together researchers and application experts interested in solving data analysis problems using combinations of data mining and visualization techniques. Researchers, software developers, and experts whose previous research had spanned both data mining and visualization were targeted to participate through membership in the Program Committee, papers, presentations, and software demos.

3.3.4 Awards

Table 2. Awards associated with the ParaText LDRD project.

Date(s)	Description	Personnel
6/08	SNL Award for Excellence: Outstanding Technical Contributions in Text Analysis	Daniel Dunlavy
8/08	SNL Award for Excellence: Organizing the 2008 Sandia Workshop on Data Mining and Data Analysis	Daniel Dunlavy
4/09	Technical Advance: SD# 11376, "Language-independent unsupervised phrase extraction from text".	Peter Chew
8/09	SNL Award for Excellence: Organizing the MAPD (Mathematics for Petascale Data Analysis) Working Group	Daniel Dunlavy

3.3.5 Students

Several students were hired as student interns as part of the ParaText LDRD project. Below are descriptions of the projects for each of those students.

3.3.5.1 Sean Gilpin

Sean Gilpin, a M.Sc. student at San Jose State University and Ph.D. student at the University of California, Davis, spent two summers at Sandia National Laboratories (2008 and 2009) working on heterogeneous ensemble classification problems. Sean was the primary architect of the HEMLOCK software framework (see Section 2.5), which he used to study the performance of heterogeneous ensemble classifier models on a variety of data analysis problems, including the classification of document collections. Two reports detailing the work conducted by Sean are available [GiDu08, GiDu09].

3.3.5.2 Taylor (Tad) Turpen

Taylor (Tad) Turpen, an undergraduate student at the University of San Diego, spent the summer of 2009 working as a student intern at Sandia National Laboratories. Tad developed a software tool for semi-supervised learning, SUNER, specifically aimed at the problem of named entity recognition (NER) from text documents. Much work has been done by the machine learning and data analysis research communities on the NER problem, with several successful NER modeling methods based on supervised learning being developed. However, to reach high levels or performance of such models, a large amount of ground truth (i.e., manually annotated, human labeled, etc.) data is required. Tad's work focused on semi-supervised learning approaches, which require only a small amount of ground truth to create models. In his work, Tad demonstrated that semi-supervised methods can be used to generate accurate NER models using much less data than that required for supervised methods. His SUNER software leverages the Stanford NER software system, which is based on conditional random field models. A report detailing the work conducted by Tad is available [TuDu09].

3.3.5.3 Becca Simon

Becca Simon, an undergraduate student at the College of Saint Benedict, is currently working as an intern at Sandia National Laboratories during the summer of 2010. Becca is working on developing spectral analysis capabilities in Trilinos for eventual use in Titan to support data clustering applications.

3.4 Publications and Presentations

The following publications and presentations were supported in part by this project.

Publications

Refereed Journal Articles

- P.A. Chew, B.W. Bader, S. Helmreich, A. Abdelali and S.J. Verzi. An Information-Theoretic, Vector-Space-Model Approach to Cross-Language Information Retrieval. *Journal of Natural Language Engineering*, to appear.

Refereed Conference Proceedings

- D.M. Dunlavy, T.M. Shead and E.T. Stanton. ParaText: Scalable Text Modeling and Analysis. In *Proc. HPDC2010: 2010 ACM International Symposium on High Performance Distributed Computing*, Chicago, IL, June 2010.
- P.J. Crossno, D.M. Dunlavy, T.M. Shead. LSAView: A Tool for Visual Exploration of Latent Semantic Modeling. In *Proc. IEEE Symposium on Visual Analytics Science and Technology*, Atlantic City, NJ, October 11-13, 2009.
- P.A. Chew, B.W. Bader and A. Rozovskaya. Using DEDICOM for Completely Unsupervised Part-of-Speech Tagging. In *Proc. NAACL-HLT Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, pp. 54-62, 2009.
- P.A. Chew. Unsupervised Phrase (and Named Entity) Extraction Improves the Results of Information Retrieval, Submitted Sept. 2009 to the 32nd European Conference on Information Retrieval.
- S.A. Gilpin and D.M. Dunlavy. Relationships Between Accuracy and Diversity in Heterogeneous Ensemble Classifiers, SIAM International Conference on Data Mining, submitted September 2009.

Book Chapters

- P.A. Chew and B.W. Bader. Algebraic techniques for multilingual document clustering. In Berry and Kogan (eds.), *Text Mining: Applications and Theory*. Chichester, UK: Wiley, 2010.

SAND Reports

- T.P. Turpen and D.M. Dunlavy. Semisupervised Named Entity Recognition. In *CSRI Summer Proceedings*, SAND2010-3083P, Sandia National Laboratories, December 2009.
- P.A. Chew, B.W. Bader and A. Rozovskaya. Using DEDICOM for Completely Unsupervised Part-of-Speech Tagging. Technical Report SAND2009-0842, Sandia National Laboratories, February 2009.
- D.M. Dunlavy, B.A. Hendrickson and T.G. Kolda, Mathematical Challenges in Cybersecurity, Technical Report SAND2009-0805, Sandia National Laboratories, February 2009.

- S.A. Gilpin and D.M. Dunlavy, Heterogeneous Ensemble Classifiers. In *CSRI Summer Proceedings*, SAND2008-8257P, Sandia National Laboratories, December 2008.
- R.A. Bartlett, D.M. Dunlavy, E.J. Guillen, T.M. Shead, J. Willenbring. Trilinos CMake Evaluation. Technical Report SAND2008-7593, Sandia National Laboratories, October 2008.
- J.D. Basilico, D.M. Dunlavy, S.J. Verzi, T.L. Bauer, W. Shaneyfelt, Yucca Mountain Licensing Support Network Archive Assistant, Technical Report SAND2008-1622, Sandia National Laboratories, September 2008.
- J.M. Brandt, D.M. Dunlavy, and A.C. Gentile, Proceedings of the 2008 Sandia Workshop on Data Mining and Data Analysis, Technical Report SAND2008-6109, Sandia National Laboratories, September 2008.

Presentations

- ParaText: Scalable Text Analysis and Visualization, SIAM Annual Meeting, Pittsburgh, PA, July 2010.
- ParaText: Scalable Text Modeling and Analysis, HPDC2010: 2010 ACM International Symposium on High Performance Distributed Computing, Chicago, IL, June 2010.
- ParaText: Scalable Text Analysis and Visualization, SIAM Conference on Parallel Processing for Scientific Computing, Seattle, WA, February 2010.
- Scalable Solutions for Data Analysis and Information Visualization, Supercomputing 09, Portland, OR, November, 2009.
- NGC Integration and HPC Analysis, presentation incorporating ParaText to the Sandia Network Grand Challenge External Advisory Board, September 22, 2009.
- ParaText: Scalable Solutions for Processing and Searching Very Large Document Collections, Sandia LDRD Day Symposium, Albuquerque, NM, September 14, 2009.
- Presentations incorporating ParaText to potential WFO customers including ODNI, SOCPAC, SOCOM JIATF and SOCOM IATF, September, 2009.
- Persistent Homology for Parameter Sensitivity in Large-scale Text-analysis (Informatics) Graphs, CSRI Workshop on Combinatorial Algebraic Topology (CAT), Santa Fe, NM, August 28-30, 2009.
- ParaText: Scalable Solutions for Processing and Searching Very Large Document Collections, NNSA LDRD 2009 Tri-Lab Symposium: Strengthening America's Infrastructure Security, Wasington, D.C., August 19, 2009.
- ParaText: Leveraging Scalable Scientific Computing Capabilities for Large-Scale Text Analysis and Visualization, 2009 SIAM Conference on Computational Science and Engineering, Miami, FL, March 2-6, 2009.
- Information Visualization with VTK, IEEE Visualization 2008, Columbus, OH, October 19, 2008.
- Text Analysis, Sandia Computer and Information Sciences Research Foundation External Review Panel Meeting, May 20-22, 2009.
- Timeline Treemap Browser, briefing to the National Technical Nuclear Forensics Center (NTNFC), part of the Domestic Nuclear Detection Office (DNDO), April 2009.
- Text Analysis and Machine Learning, presentation to potential WFO customers, February 2009.

- Coupling Informatics Algorithm Development and Visual Analysis, SIAM Annual Meeting, San Diego, CA, July 7-11, 2008.
- Heterogeneous Ensemble Classification, 2008 Sandia Workshop on Data Mining and Data Analysis, Sandia National Laboratories, Albuquerque, NM, July 22, 2008.
- Flexible Data Analysis and Visualization with Titan, 2008 Sandia Workshop on Data Mining and Data Analysis, Sandia National Laboratories, Albuquerque, NM, July 22, 2008.
- Using Visualization for Relevancy Feedback Tuning of Text Analysis Algorithms, GFX Cafe, University of New Mexico, Albuquerque, NM, April 4, 2008.
- Using Visualization for Relevancy Feedback Tuning of Text Analysis Algorithms, CSRI Seminar, CSRI, Albuquerque, NM, April 17.
- An Attempt at Group Belief Characterization and Detection, Ideology Workshop, Sandia National Laboratories, Albuquerque, NM, July 23, 2008.
- Leveraging Trilinos for Data Mining and Data Analysis, Trilinos User Group Meeting, Albuquerque, NM, November 6-8, 2007.
- Building Trilinos Using CMake, Trilinos User Group Meeting, Albuquerque, NM, November 6-8, 2007.

4. REFERENCES

[BeDuOb] Berry, Michael W., Susan T. Dumais, and Gavin W. O'Brien. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review*, 37(4):573–595, 1995.

[BeHeLeTh] C. G. Baker, U. L. Hetmaniuk, R. B. Lehoucq, and H. K. Thornquist. Anasazi software for the numerical solution of large-scale eigenvalue problems. ACM TOMS, 36(3):13:1–13:23, 2009.

[BiWa] Bian, S. and W. Wang, On Diversity and Accuracy of Homogeneous and Heterogeneous Ensembles, *Intl. J. Hybrid Intel. Sys.*, 4:103–128, 2007.

[BrDuGe] Brandt, James M. Daniel M. Dunlavy and Ann C. Gentile, Proceedings of the 2008 Sandia Workshop on Data Mining and Data Analysis, Technical report SAND2008-6109, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, September 2008.

[ChBaAb] Chew, Peter, Brett Bader and Ahmed Abdelali. Latent morpho-semantic analysis: multilingual information retrieval with character n-grams and mutual information. In *Proceedings of 22nd International Conference on Computational Linguistics (COLING 2008),* pp. 129-136, 2008.

[CrDuSh] Crossno, Patricia J., Daniel M. Dunlavy, and Timothy M. Shead, LSAView: A Tool for Visual Exploration of Latent Semantic Modeling, *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*, pp. 83-90, October 2009.

[CrHu] Crossno, Patricia J., and Warren Hunt, Timeline Treemap Browser V1.0: Temporal Analysis of Materials, Revised, Technical report SAND2009-6679, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, October 2009.

[DeDuFuLaHa] Deerwester, Scott C., Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard A. Harshman, Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[DuFuLaDeHa] Dumais, Susan T., George W. Furnas, Thomas K. Landauer, Scott Deerwester and Richard A. Harshman, Using latent semantic analysis to improve access to textual information. In *CHI '88: Proc. SIGCHI Conference on Fuman Factors in Computing Systems*, pp. 281–285. ACM Press, 1988.

[EgLoBi] Egner, M.T. and Lorch, M. and Biddle, E. Uima grid: Distributed large-scale text analysis. In Proc. of the 7th IEEE International Symposium on Cluster Computing and the Grid, pages 317–326, Washington, DC, USA, 2007. IEEE Computer Society.

[FiTa] Fielding, R.T. and Taylor, R.N. Principled design of the modern web architecture. ACM TOIT, 2(2):115–150, 2002.

[GiDu08] Gilpin, Sean A. and Daniel M. Dunlavy, Heterogeneous Ensemble Classification, in *2008 CSRI Summer Proceedings*, Technical report SAND2007-7977, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, 2008.

[GiDu09] Gilpin, Sean A. and Daniel M. Dunlavy, Relationships Between Accuracy and Diversity in Heterogeneous Ensemble Classifiers, in *2009 CSRI Summer Proceedings*, Technical report SAND2010-3083P, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, 2009.

[Go] Goldsmith, John. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198, 2001.

[LaLaDe] Landauer, Thomas K., Darrell Laham and Marcia Derr. From Paragraph to Graph: Latent Semantic Analysis for Information Visualization. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5214–5219, 2004.

[LaMcDeKi] Landauer, Thomas K., Danielle S. Mcnamara, Simon Dennis and Walter Kintsch, editors, Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, 2007.

[MaHo] Martin, Ken and Bill Hoffman. Mastering CMake, 4th Edition. Kitware, Inc., 2008.

[PlDe] Plimpton, S. and Devine, K. MapReduce-MPI Library. http://www.sandia.gov/~sjplimp/mapreduce.html.

[RaFaTuWi] Raybourn, Elaine M., Nathan Fabian, Eilish Tucker and Matthew Willis. Beyond game effectiveness part II: A qualitative study of multi-role experiential learning. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), 2010.*

[TuDu] Turpen, Taylor P. and Daniel M. Dunlavy, Semisupervised Named Entity Recognition, in *2009 CSRI Summer Proceedings*, Technical report SAND2010-3083P, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, 2009.

[Un] The Unicode Consortium. The Unicode Standard, Version 5.0 (5th Edition). Addison-Wesley Professional, 2006.

[Vi] Vigna, S. Distributed, large-scale latent semantic analysis by index interpolation. In Proc. InfoScale, pages 1–10, 2008.

[WoBoKe] Woods, K., K. Bowyer, and W. P. Kegelmeyer, Combination of Multiple Classifiers using Local Accuracy Estimates. *IEEE Trans. Pat. Recog. Mach. Int.*, 19:405–410, 1997.

[WaPaEt] Wang, W., D. Partridge, and J. Etherington, Hybrid Ensembles and Coincident Failure Diversity, in *Proc. International Joint Conference on Neural Networks*, 2001.

[WyBa] Wylie, B. and Baumes, J. A unified toolkit for information and scientific visualization. In SPIE, 2009.

DISTRIBUTION

1	MS0899	Technical Library	9536 (electronic copy)
1	MS0123	D. Chavez, LDRD Office	1011

This page intentionally left blank.

Sandia National Laboratories