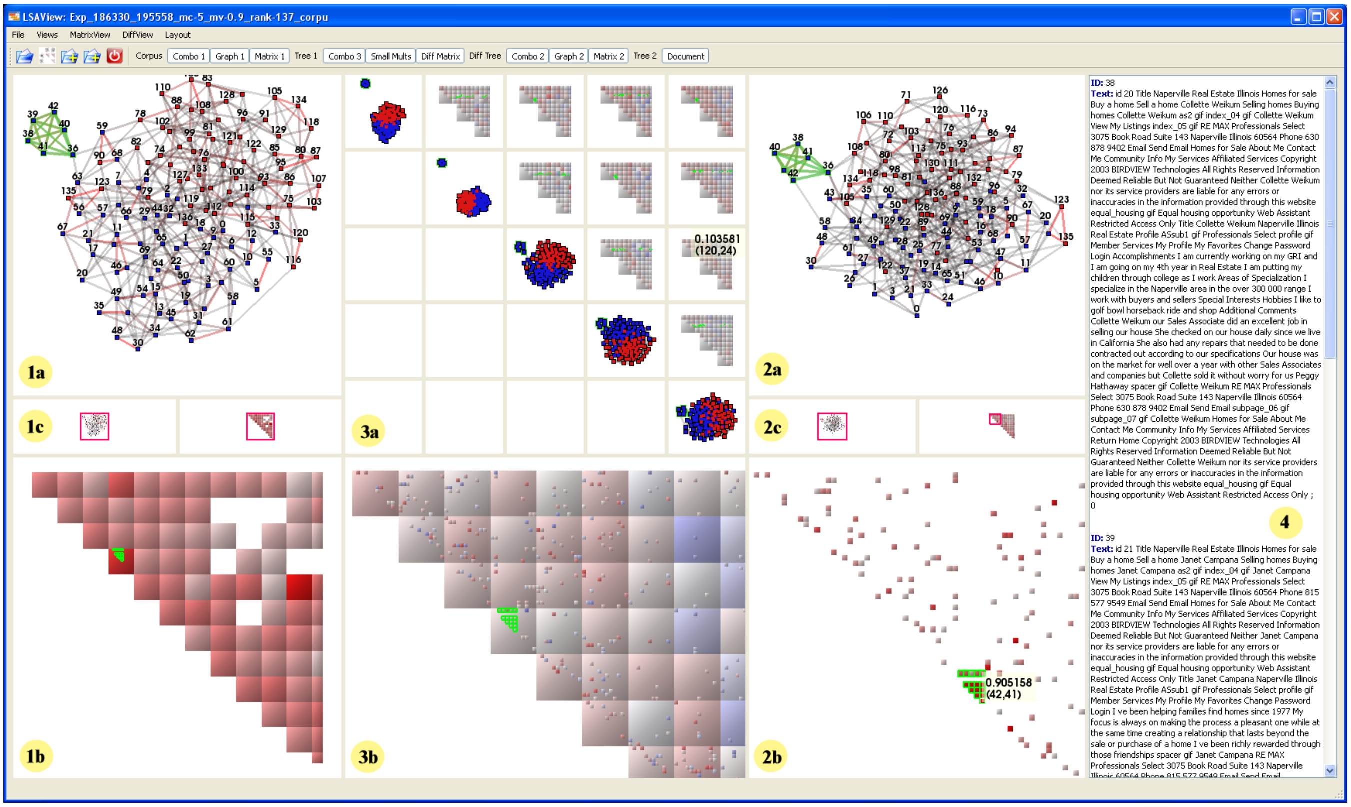
SAND2009-1976C

LSAView: A Tool for Visual Exploration of Latent Semantic Modeling

Patricia J. Crossno[*] Daniel M. Dunlavy[†] Timothy M. Shead[‡]

Sandia National Laboratories



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 ;
0

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

Figure 1: 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 (4) DOCUMENT VIEW. The CORPUS VIEW and TABLE VIEW are not shown here.

ABSTRACT

Latent Semantic Analysis (LSA) is a method used in the automated processing, modeling, and analysis of unstructured text data. One of the biggest challenges in using LSA is determining the appropriate model parameters to use for different data domains and types of analyses. Although methods have been developed to determine effective parameter choices with respect to noise in the data, our work focuses on how those choices impact analysis and problem solving. Similarly, approaches have been developed for choosing appropriate LSA model scaling parameters for information retrieval applications, but no tools currently exist for exploring the relationships between the LSA model and analysis methods. In this paper, we present LSAView, a system for interactively exploring parameter choices for LSA models. Specifically, we illustrate the use of LSAView's small multiple views, linked matrix-graph views, and data views for analysis of model selection and model application associated with graph visualization and layout.

Index Terms: I.3.8 [Computing Methodologies]: Computer Graphics—Applications; I.2.7 [Computing Methodologies]: Natural Language Processing—Text analysis

1 INTRODUCTION

Automated processing, modeling, and analysis of unstructured text (e.g., news documents, web content, journal articles, etc.) is a key task in many data analysis and decision making applications. For many applications, documents are modeled as term, or feature, vectors and latent semantic analysis (LSA) [4] is used to help model latent, or hidden, relationships between documents and terms appearing in those documents. LSA facilitates both conceptual organization and analysis of document collections as well as dimensionality reduction of the extremely high-dimensional feature vectors. In this paper we concentrate on the impact of LSA on the tasks of modeling relationships between documents using graph layout and visual clustering methods.

[*] e-mail: pjcross@sandia.gov
[†] e-mail: dmdunla@sandia.gov
[‡] e-mail: tshead@sandia.gov

LSA consists of computing a truncated singular value decomposition (SVD) of a term-by-document matrix, i.e., the collection of feature vectors associated with the documents in a text collection, or corpus. More specifically, for a term-document matrix, $A \in \mathbb{R}^{m \times n}$, its rank-$k$ SVD, A_k, is defined as

$$A_k = U_k \Sigma_k V^{\mathsf{T}} , \qquad (1)$$

where $U_k \in \mathbb{R}^{m \times k}$, $\Sigma_k \in \mathbb{R}^{k \times k}$, $V_k \in \mathbb{R}^{n \times k}$ contain the k leading left singular vectors, singular values, and right singular vectors, respectively. Furthermore, $U_k^{\mathsf{T}} U_k = V_k^{\mathsf{T}} V_k = I_k$, 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 both a reduction in model noise and computation required in many analysis methods.

A central challenge in using LSA for text analysis is determining appropriate parameters for computing and applying the SVD; specifically, selecting the rank of the SVD and scaling of the singular values for different data and types of analyses. The rank selection problem refers to the determination of an appropriate rank of the truncated SVD for use with a particular task and data set. For the problem of document clustering, rank selection is typically performed by analyzing document sets related to the collection to be clustered to determine a suitable rank. Clusterings for these related collections are used to tune the LSA rank parameter for the collection to be clustered [10]. That approach requires annotated document collections whose term-document relationship distributions are highly correlated with those of the document collection to be clustered. Such annotations may be laborious to generate and the results may contain errors or subjective clusterings. Furthermore, determining how closely related the underlying distributions of term and document relationships between two document collections is not well understood. Thus, a new technique for solving the rank selection problem is needed for the problem of document clustering.

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 rank-k SVD. In this work, we use cosine similarities, defined as

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

between documents i and j, where $\langle \cdot, \cdot \rangle$ is the standard inner product and v_k^i is the ith row of V_k. 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 the weighted adjacency matrix for constructing 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.

Singular value scaling (or rescaling) refers to an exponential scaling of Σ_k by a scalar $\alpha/2 \in \mathbb{R}$. The result of singular value scaling is a contraction ($0 < \alpha < 2$), expanion ($\alpha > 2$), inversion ($\alpha < 0$) or flattening ($\alpha = 0$) the singluar value spectrum. Such scaling is motivated by different applications; for example, in document clustering, inverting the singular values tends to highlight more of the novel or anamolous relationships between documents. For the document clustering problem, the use of singular value scaling changes the similarity scores in (2) to

$$e_{ij}(k,\alpha) = \frac{\langle v_k^i \Sigma_k^{\alpha/2}, v_k^j \Sigma_k^{\alpha/2} \rangle}{\|v_k^i \Sigma_k^{\alpha/2}\| \|v_k^j \Sigma_k^{\alpha/2}\|} . \qquad (3)$$

Although originally developed to improve information retrieval systems [1, 17], singular value scaling can be used for any analysis task employing LSA models. However, as for the rank selection problem, no tools exist for visually exploring the relationships between the scaling parameter and analysis methods.

In this paper we present LSAView, a system for interactively exploring the impact of parameter choices in informatics analysis pipeline systems on the visual presentation and analysis capabilities that data analysts utilize in decision making processes. Specifically, we present the visualization capabilities of LSAView and illustrate how they can be used for understanding the relationships between parameters used in LSA and graph-based cluster analysis. LSAView fills a gap for algorithms developers who require better understanding of the impact and sensistivities of parameters in their methods and for data analysts who need to better understand the models used in their analyses. Through visual exploration both developers and analysts can explore the complex relationships between algorithms, models and analysis.

The major contributions of this works as as follows:

- A framework for visually exploring the relationship between LSA model parameters and graph clustering methods.
- A new matrix view to support scalable, zoomable visualization of matrix and matrix difference data.
- Visualization of graph statistics for identifying unexpected edges associated with LSA model parameters.
- Case studies illustrating the use of visual algorithm analysis for identifying the impact of LSA model parameters on graph layout and clustering methods.

The remainder of this paper is organized as follows. In Section 2, we discuss related work in the areas of LSA and visualization. LSAView is described in Section 3. Section 4 illustrates the use of LSAView in two case studies, and conlusions are presented in Section 5.

2 Related Work

2.1 LSA: Rank Selection and Singular Value Scaling

In the case of rank selection, methods exist based on a variety of techniques: for example, cross validation [13, 15], Bayesian model selection via Markov chain Monte Carlo methods [8], expectation maximization [14], and Bayesian inference [11]. A recent survey also indicates challenges associated with rank selection for the problem of LSA [2]. However, such methods focus on rank determination with respect to noise in the the data and not with respect to how the SVD will be used for analysis, such as document clustering. Attempts at determination of a useful SVD rank on particular problems exist as well [9], but tools for general exploration do not yet exist for understanding the relationships between the rank of the SVD and its impact on text analysis methods.

Similarly, methods for scaling the singular values have been developed for information retrieval applications [1, 17], but no tools for exploring the relationships between spectral properties and analysis methods currently exist.

2.2 Visualization

Eick et al developed a multi-view system for evaluating supervised computational linguistics algorithms [5]. After training the system, they measured classification accuracy against a set of predefined categories and concepts. Outside of text analysis, Groth described a multi-view system to examine the performance of a naive Bayes classifier with respect to various discretization schemes [7].

3 LSAView

LSAView is built using VTK's Titan Informatics Toolkit. [16] [3] It uses a multiple-coordinated view approach to exploring the impact of parameter choices, as shown in Figure 1. The application's input data consists of a corpus with the text for each document, and one or more document similarity graphs, each produced using a different parameter value, such as rank.

The application's views are designed to provide both alternative representations of a single result, and to provide comparative views of multiple results. The graph views show changes in clustering, while the matrix views use color-coding to permit visual comparisons of edge weights, statistics calculated over a range of result graphs (see Section 3.2.1), or explicit differences in values or statistics between two graphs. Non-graphical views enable drill-down to explicit numerical values for the edge weights and statistics for each edge. Text-based views include a corpus document list and a document viewer to enable examination of source texts for selected documents.

The graphical views are grouped into three panels. In Figure 1, these panels are enabled along with the document viewer on the far right. The left and right graphical panels provide detailed inspection of two document similarity graphs resulting from different rank choices. The upper graph view and lower matrix view each represent the same data in a different form. In between them, are two you-are-here views, which provide both context and an alternative navigation method within the corresponding window.

The middle panel contains two views, a small multiples view at the top and a differences matrix at the bottom. The differences matrix shows color-coded differences between the left and right panels' edge weights. The small-multiples view provides a high level view of five different rank results, though not necessarily the same as the ones that are shown in the side panels. To assist in managing the large number of views, a series of display toggle buttons for both individual views and for entire panels is provided in the toolbar at the top of the application. Additionally, double-clicking within a view will expand it to fill the application window, simultaneously turning off all other views. Double-clicking again will restore the previous view configuration. For graph or matrix views with an associated you-are-here view, double-clicking includes the you-are-here view in the expanded view.

3.1 Graph Views

The graph views display the document similarity matrix as a node-link diagram, where nodes represent documents and edges are the weighted similarities between them. The graph provides a high-level view of how document clusters change relative to parameter changes. In the small-multiples view in Figure 1, the five graph views show the impact of changing rank on the connectivity of a document group. Since the choice of graph layout algorithm can impact the perceived groupings, LSAView provides interactive user selection from several different layout algorithms found in Titan's graph layout strategies.

Nodes are labeled with the document number and colored by category, when category labels are available. Although in real world applications the correct categorization is typically not known, for sample problem sets that have been hand classified, node coloring allows comparison between the correct groupings and those produced by the algorithm. In Figure 1, the corpus has been partitioned into two categories and we are examining the disjoint cluster of blue nodes, which are separated from the main group.

Edges are color-coded using increasing saturation to denote increasing similarity values, with low values in gray and high values in red. Nodes, edges, or edges and nodes contained in a rectangular region can be selected. Selections are highlighted in green and are linked to all other views. Note that not all views share the same edges, so each view will be limited to highlighting only selected edges that they possess.

3.2 Matrix Views

The matrix views display each edge of the similarity graph's sparse matrix as a rectangle. To provide a scalable matrix representation, we transform the edge table into a tree. Although the actual edges form a sparse matrix, the tree is constructed as though there were a dense matrix formed by a complete graph between the documents. At each level of the tree, the square root of the number of children in the complete graph is used to create a set of bins (plus additional bins in each dimension if there is a remainder) into which the existing children are partitioned. The tree is built from the leaves up, so regions of the matrix that do not have any edges are left empty.

The initial rendering of the matrix view displays the entire tree at the lowest level that preserves a minimum rectangle size. This facilitates the matrix view's use in meta-views, such as the small-multiples view (Section 3.4). The user interface includes zooming, so all levels of the tree are accessible. As the user zooms into a deeper level, the rendering of the new level overlaps the old level to provide context. Once the size of the rectangles at the lower level exceeds a certain threshold, the higher level rectangles are no longer rendered. The differences matrix in Figure 1 demonstrates the overlapped rendering, whereas the matrix view to its right shows how the zoom will look after the threshold is crossed (note that this is the leaf level of the tree, so only the actual edges are rendered).

3.2.1 Matrix Data

We define the sample mean of $e_{ij}(k,\alpha)$ using $n+1$ samples as

$$\bar{e}_{ij}(k,\alpha,n) = \frac{1}{n+1}\sum_{r=k-n/2}^{k+n/2} e_{ij}(r,\alpha)\ , \tag{4}$$

and the corrresponding standard error as

$$s_{ij}(k,\alpha,n) = \sqrt{\frac{1}{n}\sum_{r=k-n/2}^{k+n/2}\left(e_{ij}(r,\alpha)-\bar{e}_{ij}(k,\alpha,n)\right)^2}\ , \tag{5}$$

where α is the singular value scaling parameter. Note that the statistics use biased samples centered around edge weights associated with a rank-k LSA model. The purpose of these statistics are to help identify anamalous edge weights given variances in those weights across the most closely related LSA models.

Using the sample mean and standard error definitions above, we define a one-sample t statistic with sample size of $n+1$ for the weight on the edge between nodes i and j corresponding to the rank-k SVD and singular value scaling value of α as

$$t_{ij}^{(1)} = \frac{\bar{e}_{ij}(k,\alpha,n)-e_{ij}(k,\alpha)}{s_{ij}(k,\alpha,n)/\sqrt{n+1}}\ . \tag{6}$$

This one-sample t statistics can be used to identify anomalous, or outlier, edge weights in a single graph. The hypothesis being tested is that there is no difference between an edge weight and its mean value (sampled from weights derived from different LSA models); thus, higher values of the t statistic correspond to more anomlaous edge weights.

Similarly, a two-sample t statistic for the corresponding edge weights using SVDs with ranks k_1 and k_2 and sample sizes n_1 and n_2, respectively, is defined as

$$t_{ij}^{(2)} = \frac{\bar{e}_{ij}(k_1,\alpha,n_1)-\bar{e}_{ij}(k_2,\alpha,n_2)}{\sqrt{\frac{[s_{ij}(k_1,\alpha,n_1)]^2}{n_1}+\frac{[s_{ij}(k_2,\alpha,n_2)]^2}{n_2}}}\ . \tag{7}$$

This two-sample t statistics is used to identify anomalous edge weights when comparing two graphs. The hypothesis being tested here is that the means weights of corresponding edges in the two graphs are not different.

As the similarities defined in (2) form the entries of a similarity matrix, E, the t statistics defined in (6) and (7) form the entries of the matrices, $T^{(1)}$ and $T^{(2)}$, respectively. Thus, these statistics can be viewed for entire graphs using the matrix views defined above.

For all statistics defined above for sparse similarity matrices (see Section 1), edge weights of 0 are treated as missing values and the sample sizes are adjusted to reflect this.

3.2.2 Visualization

A matrix view associated with a single document similarity graph can be used to visualize edge weights or one-sample t statistics. When viewing edge weights, different options are available for propagating values to higher level nodes in the tree. The choices are to take either the minimum, maximum, or average of the children's weights as a summary edge weight value for a non-leaf node. For the t statistics, the maximum value for each node's children is stored as its value.

For difference matrix views, a set of variables derived from a pair of similarity graphs is provided for each edge. Other than the two-sample t statistics, all other choices are calculated by subtracting the edge weights, sample means, or standard errors of the second matrix from the first. In all difference matrix views, the maximum value is propagated upward.

Edge and node rectangles are color-coded to permit visual comparisons between graphs. All of the values to be color-coded, except for the t statistics, range from negative one to positive one. Saturation is used to show increasing absolute value, with zero encoded as white. Large positive values are shown in bright red, large negative values in deep blue. The lookup table is constructed using a linear ramp. Selected edge rectangles are outlined in green to stand out against the blue-red palette and selections are linked to all other views.

However, for the t statistics, values are only limited to be being non-negative numbers. So we change hue to show that meaning of the encoding is different and use saturation ranging from white to bright green to encode increasing value. The lookup table is constructed using a log scale to focus color and attention on the highest values in the matrix. Selected edge rectangles are outlined in red to stand out against the green palette and selections are linked to all other views.

3.3 You-Are-Here Views

The you-are-here views provide context for both graph and matrix views so that the user can keep track of their location within the high-level view as they zoom-in to focus on a small region. The red rectangle in the you-are-here view shows the current view boundaries and position within the larger graph or matrix view. The you-are-here view is implemented simply as a rectangle drawn over a captured image. Updates to the image are only made when the color-coded variable changes, so selections are not visible in the view.

Panning or zooming in the graph or matrix view will update the rectangle location and size. Similarly, dragging or scaling the red rectangle will pan or zoom the contents of the graph or matrix view. Often navigating using the you-are-here view is preferred because of the contextual landmarks it provides.

3.4 Small Multiples Views

The small-multiples view is a combination of graph and matrix views that enables comparisons of up to five different document similarity graphs. The graph views provide a high-level overview of the clustering impacts resulting from the parameter changes . Additionally, the graph views act as labels to define the matrix pairs that were used to calculate the difference matrices. Each difference matrix represents the difference between the graph at the head of its row and the graph at the tail of its column.

Each of the graph and matrix views is fully interactive and operates just as any graph or matrix view elsewhere in the application would. Selections made in any one of the views are fully linked to all other views in the application. Zooming and panning permit exploratory navigation within each view. The only limitation is the lack of you-are-here views for each graph or matrix, so it is difficult to maintain context. Another difference is that double-click expands the entire small multiples view, rather than any one view within it.

4 Case Studies

Two case studies are presented in this section, focusing on the problems of rank selection and singular value scaling, respectively. These studies illustrate how LSAView can be used to interactively determine suitable parameters for LSA models for the problem of graph clustering.

4.1 Data

Two sets of data are used in the case studies. The first set, denoted DUC, consist of newswire documents used in the 2003 Document Understanding Conference (DUC) for evaluating summarization systems on clusters of documents [12]. The DUC data is comprised of 298 documents in 30 clusters, with each cluster containing about 10 documents focused on a particular topic or event. This data was manually annotated and thus reflects human judgement with respect to cluster assignment of the documents.

The second set of documents, denoted TECHTC, is from the TechTC-100 Test Collection[1] [6]. Each of the 100 subsets of documents in this collection consists of about 150 HTML documents partitioned into two clusters. These data sets were used to evaluate binary classification models, but work well in evaluating clustering algorithms as well. An important feature of these data sets is that identified with each set is a difficulty rating associated with determining the clusterings. These ratings were computed using a variety of supervised learning classification models and are thus only an approximation of the true difficulty of solving the associated clustering problem. The case study results presented here use the `Exp_186330_195558` subset of the TECHTC data. This particular subset is the one with the lowest difficulty rating. Using data that is easy to cluster will hopefully illustrate the strength of LSAView to highlight subtle differences in the LSA models.

4.2 Rank Selection

The process of using LSAView to visually determine the rank of the LSA model most suitable for graph clustering is as follows:

1. Identify a range of potential ranks using the small multiples view.
2. Choose a suitable rank by comparing graph clusterings using the graph, matrix, and data table views.
3. Validate the chosen rank using the document view.

Note that at each step of this process, several iterations may be required.the user.

During step 1 of the rank selction process, coarse steps in rank values can be used to identify changes in the graph clustering of documents over a wide range of LSA model ranks. Figure 2 illustrates this for the DUC data using LSA model ranks of $k = 10, 30, 50, 70, 90$. The difference matrix views in the figure are colored by differences in edge weights (i.e., document similarities). By

[1] http://techtc.cs.technion.ac.il/techtc100/techtc100.html

visualizing the impact of the rank on both the graph clustering and the changes in the edge weights simultaneously over a wide range of LSA model ranks, the range of suitable ranks can be narrowed. Specifically, in Figure 2, the LSA model ranks of $k = 10$ and $k = 30$ appear to have both good separations of document groups and edge weights that are somewhat differentiated from those associated with the other ranks (as depicted by the bold blue and red edge rectangles). Thus subsequent investigation should focus on ranks closer to those values.

After some iteration of step 1, we arrive at Figure3, depicting LSA models ranks of $k = 28, 29, 30, 31, 32$. In this figure, we have switched the difference matrix views to be colored by the sample means of the edge weights. As the ranks are so close, these difference matrix views better help to identify the changes in the edge weights, as the means of those weights should also be close. Note that the two-sample t statistics could be viewed as well for similar analysis.

We now move on to step 2 in the rank selection process, depicted in Figure 4 where graph views and matrix views colored by two-sample t statistics are shown for DUC data with ranks of $k = 30$ (left) and $k = 32$ (right). The t statistics can be used to quickly identify differences between the two graphs, and a user is easily drawn to the areas with the most significant differences between edges weights (i.e., bold green edge and node rectangles in the difference matrix view).

After zoomed inspection of graph clusterings associated with several of the most significant differences, we arrive at Figure 5, illustrating anomalous links (determined from the t statistics) between document 297 and groups of highly related documents (i.e., linked by bold red links). Now the similarities of document 297 with other documents differs dramatically for the LSA models of rank $k = 30$ (left) and $k = 32$ (right). Thus, further inspection of that document is required to help identify the most suitable rank.

This brings us to step 3 of the rank selection process, where manual inspection of the underlying documents is used to validate the selected rank. Figure 6 presents the document view aside the zoomed view of the LSA model of rank $k = 30$. After reading the document and those identified as most similar (i.e., those linked to document 297 in the graphs) for the different LSA model ranks (including rank $k = 32$ and other nearby ranks), we conclude that a rank of $k = 30$ is most suitable. Note that the fact that the most suitable rank is equal to the number of underlying clusters is coincidental in this case.

Document 297, about China's leader's statements and policies regarding separatists, turns out to be an anomaly, in that it is only tangentially related to the documents in the cluster to which it is assigned where the main topic is specifically the trial of 3 separatists in China. It appears to be more related to another group of documents, documents 150–158, about the policies and responses of the Russian government to Chechnyan separatists. Such subtle relationships would be difficult to assess by simply reading all of the documents. By orgaininizing and modeling the data using LSA, combined with interactive exploration of different LSA models, we were able to determine such sublte relationships by reading only a few key documents. The strength of such visual algorithm analysis can aid both the developer and analyst in their understanding of the LSA modeling process.

4.3 Singular Value Scaling

The process of using LSAView to visually determine the singular value scaling for the LSA model that is most suitable for graph clustering follows the same general steps as for the rank selection problem. Again, we use the small multiples view to determine general trends of the impact of the LSA model parameters—in this case, the singular value scaling parameter, α—followed by by more details visual analysis using the graph and matrix views, and finally

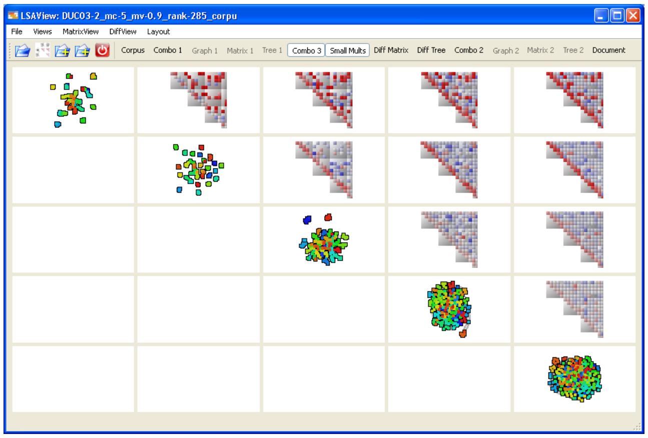


Figure 2: Small multiples view of DUC data with LSA model ranks of $k = 10, 30, 50, 70, 90$. The difference matrix views depict differences in the edge weights across the different graphs.

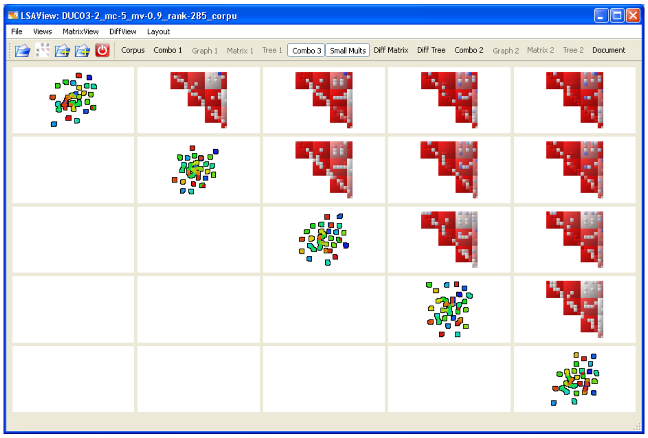


Figure 3: Small multiples view of DUC data with LSA model ranks of $k = 28, 29, 30, 31, 32$. The difference matrix views depict differences in the sample means of the edge weights across the different graphs.

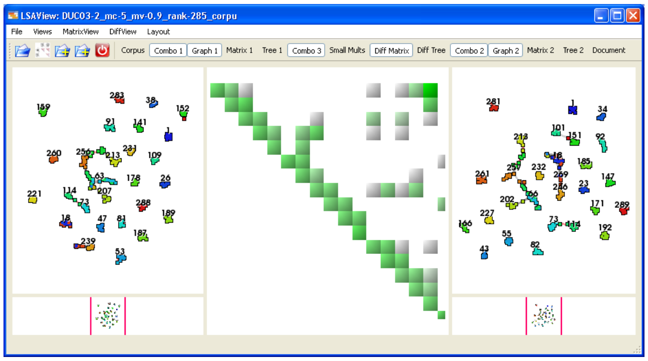


Figure 4: Graph model comparisons of DUC data with rank $k = 30$ (left) and $k = 32$ (right) using the difference matrix view (center) of two-sample t statistics.

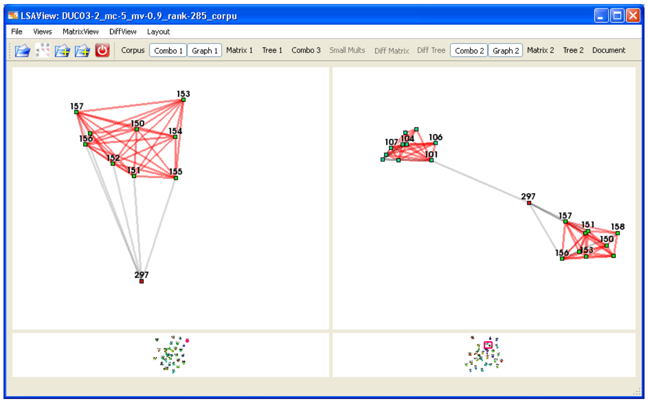
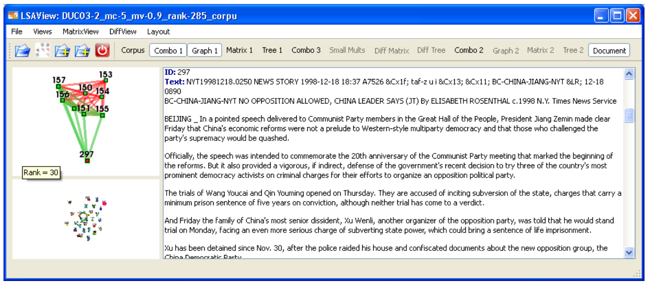

LSAView: DUC03-2_mc-5_mv-0.9_rank-285_corpu
File Views MatrixView DiffView Layout
Corpus Combo 1 Graph 1 Matrix 1 Tree 1 Combo 3 Small Mults Diff Matrix Diff Tree Combo 2 Graph 2 Matrix 2 Tree 2 Document
157
153
150
154
156
152
151
155
297
107
104
106
101
297
157
151
158
150
156
153

Figure 5: Graph model comparisons of DUC data with rank $k = 30$ (left) and $k = 32$ (right) depicting weak document similarities (grey edges) between node 297 and groups of highly related documents (red edges).


LSAView: DUC03-2_mc-5_mv-0.9_rank-285_corpu
File Views MatrixView DiffView Layout
Corpus Combo 1 Graph 1 Matrix 1 Tree 1 Combo 3 Small Mults Diff Matrix Diff Tree Combo 2 Graph 2 Matrix 2 Tree 2 Document
157
153
156
150
154
151
155
297
Rank = 30

ID: 297
Text: NYT19981218.0250 NEWS STORY 1998-12-18 18:37 A7526 &Cx1f; taf-z u i &Cx13; &Cx11; BC-CHINA-JIANG-NYT &LR; 12-18 0890
BC-CHINA-JIANG-NYT NO OPPOSITION ALLOWED, CHINA LEADER SAYS (JT) By ELISABETH ROSENTHAL c.1998 N.Y. Times News Service

BEIJING _ In a pointed speech delivered to Communist Party members in the Great Hall of the People, President Jiang Zemin made clear Friday that China's economic reforms were not a prelude to Western-style multiparty democracy and that those who challenged the party's supremacy would be quashed.

Officially, the speech was intended to commemorate the 20th anniversary of the Communist Party meeting that marked the beginning of the reforms. But it also provided a vigorous, if indirect, defense of the government's recent decision to try three of the country's most prominent democracy activists on criminal charges for their efforts to organize an opposition political party.

The trials of Wang Youcai and Qin Youming opened on Thursday. They are accused of inciting subversion of the state, charges that carry a minimum prison sentence of five years on conviction, although neither trial has come to a verdict.

And Friday the family of China's most senior dissident, Xu Wenli, another organizer of the opposition party, was told that he would stand trial on Monday, facing an even more serious charge of subverting state power, which could bring a sentence of life imprisonment.

Xu has been detained since Nov. 30, after the police raided his house and confiscated documents about the new opposition group, the China Democratic Party.

Figure 6: Manual inspection of documents associated with anamalous edge weights can be performed using the linked graph and document views. Interacting with both the graphs and the underlying data is necessary for determining where LSA models are linking nodes differently than an analyst expects.

validation using the document view. The main difference is that it may be useful to first inspect the scaled singular values directly to determine if one scaling may be significantly different. Figure 7 presents the scaled singular values of of the TECHTC data for $\alpha = -2, -1, -0.5, 0, 0.5, 1, 2$. Note that the original singular values are those scaled by $\alpha = 2$.

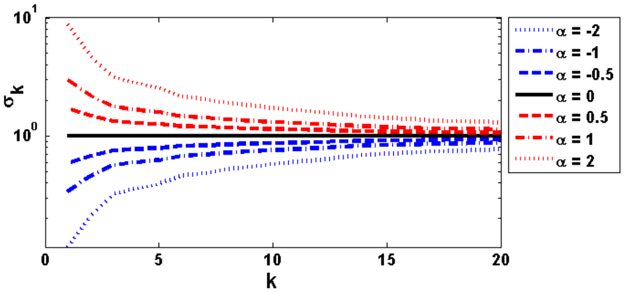


Figure 7: Singular values up to $k = 20$ for the TECHTC data scaled using different values of α. The original singular values correspond to $\alpha = 2$.

5 Conclusion

Acknowledgements

This work was funded by the Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories, 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.

References

[1] H. Bast and D. Majumdar. Why spectral retrieval works. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18, New York, NY, USA, 2005. ACM Press.

[2] R. B. Bradford. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 153–162, New York, NY, USA, 2008. ACM.

[3] P. Crossno, B. Wylie, A. Wilson, J. Greenfield, E. Stanton, T. Shead, L. Ice, K. Moreland, J. Baumes, and B. Geveci. Intelligence analysis using titan. In *Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 241–242, Nov 2007.

[4] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *CHI '88: Proceedings of the SIGCHI Conference on Fuman Factors in Computing Systems*, pages 281–285. ACM Press, 1988.

[5] S. G. Eick, J. Mauger, and A. Ratner. A visualization testbed for analyzing the performance of computational linguistics algorithms. *Information Visualization*, 6(1):64–74, 2007.

[6] E. Gabrilovich and S. Markovitch. Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5. In *Proc. International Conference on Machine Learning (ICML)*, pages 321–328, Alberta, Canada, July 2004. Banff.

[7] D. P. Groth. Visualizing distributions and classification accuracy. In *Proc. International Conference on Information Visualization*, pages 389–394, July 2006.

[8] P. D. Hoff. Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102(478):674–685, 2007.

[9] T. K. Landauer, D. Laham, and M. 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.

[10] K. Lerman. Document clustering in reduced dimension vector space. http://www.isi.edu/∼lerman/papers/Lerman99.pdf, 1999.

[11] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, November 2003.

[12] P. Over and J. Yen. An introduction to DUC-2003: Intrinsic evaluation of generic news text summarization systems. In *Proc. DUC 2003 workshop on text summarization*, 2003.

[13] A. B. Owen and P. O. Perry. Bi-cross-validation of the SVD and the non-negative matrix factorization. Technical report, Stanford University, May 2008.

[14] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001.

[15] S. Wold. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405, 1978.

[16] B. Wylie and J. Baumes. A unified toolkit for information and scientific visualization. In K. Borner and J. Park, editors, *Proc. Visualization and Data Analysis 2009*, volume 7243, page 72430H. SPIE, 2009.

[17] H. Yan, W. I. Grosky, and F. Fotouhi. Augmenting the power of LSI in text retrieval: Singular value rescaling. *Data Knowledge and Engineering*, 65(1):108–125, 2008.