International Journal on Artificial Intelligence Tools

TopicView: Visual Analysis of Topic Models and Their Impact on Document Clustering

Patricia J. Crossno, Andrew T. Wilson, Timothy M. Shead and Warren L. Davis, IV
Scalable Analysis and Visualization, Sandia National Laboratories
Albuquerque, NM 87185, USA
{*pjcross,atwilso,tshead,wldavis*}*@sandia.gov*

Daniel M. Dunlavy
Data Analysis and Informatics, Sandia National Laboratories
Albuquerque, NM 87185, USA
dmdunla@sandia.gov

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

We present a new approach for analyzing topic models using visual analytics. We have developed TopicView, an application for visually comparing and exploring multiple models of text corpora, as a prototype for this type of analysis tool. TopicView uses multiple linked views to visually analyze conceptual and topical content, document relationships identified by models, and the impact of models on the results of document clustering. As case studies, we examine models created using two standard approaches: Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Conceptual content is compared through the combination of (i) a bipartite graph matching LSA concepts with LDA topics based on the cosine similarities of model factors and (ii) a table containing the terms for each LSA concept and LDA topic listed in decreasing order of importance. Document relationships are examined through the combination of (i) side-by-side document similarity graphs, (ii) a table listing the weights for each document's contribution to each concept/topic, and (iii) a full text reader for documents selected in either of the graphs or the table. The impact of LSA and LDA models on document clustering applications is explored through similar means, using proximities between documents and cluster exemplars for graph layout edge weighting and table entries. We demonstrate the utility of TopicView's visual approach to model assessment by comparing LSA and LDA models of several example corpora.

Keywords: text analysis; visual model analysis; latent semantic analysis; latent dirichlet allocation; clustering

1. Introduction

Latent Semantic Analysis (LSA)[1] and Latent Dirichlet Allocation (LDA)[2] are two popular mathematical approaches for modeling conceptual/topical content and relationships in textual data collections. Questions posed by algorithm developers and data analysts working with these models motivated the work described in this pa-

per: How closely do LSA's concepts correspond to LDA's topics? How similar are the most significant terms in LSA concepts to the most important terms of corresponding LDA topics? Are the same documents affiliated with matching concepts and topics? Do the document similarity graphs produced by the two models identify corresponding groups of documents? How well do document groups found in their respective similarity graphs match human-generated clusters? What is the impact of these different topic models on analysis applications such as document clustering, categorization and summarization?

LSA, LDA, and other topic (or factor) models of textual data have much in common: They use bag-of-words modeling (ignoring syntactic and grammatical structure), begin by transforming text corpora into term-document frequency matrices, reduce the high dimensional term spaces of textual data to a user-defined number of dimensions, produce weighted term lists for each concept or topic, produce concept or topic content weights for each document, and produce outputs that can be used to compute document relationship measures or as inputs to other analysis applications. LSA uses a singular value decomposition (SVD) of the term-document frequency matrices (which are often transformed or scaled) to define a basis for a shared semantic vector space, in which the maximum variance across the data is captured for a fixed number of dimensions. In contrast, LDA employs a Bayesian model that treats each document as a mixture of latent underlying topics, where each topic is modeled as a mixture of word probabilities from a vocabulary. Although LSA and LDA outputs can be used in similar ways, their output values represent entirely different quantities, with different ranges and meanings. LSA produces term-concept and document-concept correlation matrices, with values ranging between -1 and 1 where negative values indicate inverse correlations. LDA produces term-topic and document-topic probability distribution matrices, where probabilities range from 0 to 1. Direct comparison and interpretation of similarities and differences between LSA and LDA models is thus an important challenge in understanding which model may be most appropriate for a given analysis task.

Our approach is to move away from statistical comparisons, focusing instead on human consumable differences that reflect how these models are typically used for analysis tasks. Although applications may use a variety of metaphors to visualize document collections, including scatter plots,[3] graphs,[4] and landscapes,[5] all of these methods rely on document similarity measures or clustering to position documents within a visualization. These representations are often combined with labels to identify the topical or conceptual content of document groups.[6] Consequently, we focus our comparison on the document relationships and conceptual categories identified by LSA and LDA models, along with the impact of the different models on the analysis tasks of computing document similarities and clustering documents.

In this paper we discuss *TopicView*, a visual analytics application designed to visually compare and interactively explore multiple topic models from this user-based perspective. In TopicView, tabbed panels of linked views compare conceptual content (i.e. *concepts* and/or *topics*), document relationships (between individual

documents and between documents and conceptual content), and clusters computed using modeled documents. In addition to describing the design and implementation of TopicView, we present insights on differences between LSA and LDA models that were revealed using TopicView with several corpora. We specifically focus on LSA and LDA models in this work to illustrate the power of a visual analysis approach, although the techniques presented here would be applicable to other topic models as well.

2. Related Work

Blei et al. discuss the differences in accuracy between LSA and LDA when these models are used with SVM classification.[2] Similarly, Kakkonen et al. compare the accuracy of these models with other techniques such as k-Nearest Neighbor to perform classification, and measure the resultant accuracy in terms of Spearman correlation.[7] Other statistics, such as F-Measure and receiver operating characteristic (ROC) curve values, have been used to measure differences between LSA and LDA models as well.[8] Although the utility of the work presented in those papers may be evident to algorithm developers and machine learning researchers, our experience with data analysts and decision makers across several domains indicates that such results are difficult to incorporate into workflows and decision making processes required in real-world data analysis activities. It is often difficult for practitioners of these models and analysis capabilities to translate statistical confidences into advice on how to choose a particular analysis method and how best to use that method for solving a variety of tasks.

To assess how well existing methods model human semantic memory, Griffiths et al. compare generative probabilistic topic models with models of semantic spaces.[9] They are concerned with a model's ability to extract the gist of a word sequence in order to disambiguate terms that have different meanings in different contexts. This is also related to predicting related concepts. LSA and LDA are used as instances of these approaches and compared in word association tasks.

In contrast to the aforementioned work, our research compares the impact that model differences have on visual analytics applications, not solely on statistical differences. We focus on how users will derive conclusions driven by visual analysis, and use the statistical measures as a supplement to our findings.

Collins et al. combine tag clouds with parallel coordinates to form Parallel Tag Clouds, an approach for comparatively visualizing differentiating words within different dimensions of a text corpus.[10] Word lists are alphabetical, with word size scaled according to word weight. Similar to parallel coordinates, matching terms are connected across columns.

Although we have similar goals in comparing term lists, we believe that our approach of sorting terms by weight, combined with scaling text luminance by weight, provides a clear comparison of the relative significance of terms across concepts and topics. Our approach avoids the layout complications and potential overlaps

encountered when words are drawn at vastly different scales.

3. Modeling and Analysis Approaches

In this section, we provide a brief overview of the models, algorithms, and graph data abstractions used in our work and discussed in this paper. More detailed descriptions of the various models and algorithms can be found in the references provided.

3.1. *Latent Semantic Analysis*

LSA was originally developed to improve indexing and retrieval accuracy in search tasks by statistically modeling relationships between terms distributed across documents (specifically synonymy and polysemy).[1] Over time, LSA models have grown to be used for other document analysis tasks, including clustering, categorization, cross-language retrieval, and summarization.

LSA computes a truncated SVD of a term-document matrix, i.e., the collection of m weighted term vectors associated with the n documents in a corpus of text.[11] More specifically, the k-dimensional LSA model of a term-document matrix, $A \in \mathbb{R}^{m \times n}$, is its rank-$k$ SVD,

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

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. The k latent features, or concepts, are linear combinations of the original terms, with weights specified in U_k. Documents are modeled as vectors in concept space, with coordinates specified in V_k.

3.2. *Latent Dirichlet Allocation*

LDA is a probabilistic generative approach that models a collection of documents using topics, which are probability distributions over a vocabulary.[2] A generative model is constructed using a vocabulary of m distinct words, a number of topics k, two smoothing parameters α and β, and a prior distribution over document lengths (typically Poisson). Such a model can be used to generate random documents whose contents are mixtures of topics.

To model the topics in an existing corpus with LDA, the parameters of the generative model must be learned from the data. Specifically, for a corpus containing n documents and m terms, the k-topic LDA model of a term-document matrix, $A \in \mathbb{R}^{m \times n}$, is

$$A_k = \phi \theta^{\mathsf{T}} \ , \tag{2}$$

where $\phi \in \mathbb{R}^{m \times k}$ is the matrix of term probabilities for each topic, and $\theta \in \mathbb{R}^{n \times k}$ contains topic probabilities for each document. The remaining parameters α, β and k are specified by the user. For the LDA models used in this paper, parameter fitting is performed using collapsed Gibbs sampling[12] to estimate θ and ϕ.

3.3. *Document Similarity Graphs*

To identify related documents, we compute the cosine similarity between all pairs of documents. For LSA models, these similarities are computed between the scaled document vectors, i.e., the rows of $V_k\Sigma_k$. For LDA models, they are computed between the rows of θ. The similarities form a similarity matrix that can be interpreted 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. To support analysis of large corpora, only edge weights above a threshold are used, leading to sparse similarity matrices. Finally, graph layout methods are used to reveal document groups by positioning related nodes near one another in the resulting graph.

3.4. *Clustering*

Although topic models are often used to identify document clusters directly from the model factors, in this paper we investigate the impact of using the LSA and LDA models as input to clustering algorithms. Specifically, we cluster the modeled documents (the rows of $V_k\Sigma_k$ for LSA and θ for LDA) using the K-means clustering algorithm.[13] K-means separates data points into groups, or clusters, surrounding a central point, called a centroid. These centroids are arranged such that the total intra-cluster dissimilarity to each centroid is minimized. Alternatively, K-means can be used with similarities by maximizing the total intra-cluster similarity.

For n observations, $X = \{x_1, .., x_n\}$, and k centroids, $C = \{c_1, .., c_k\}$, the function to be minimized in the K-means algorithm, $F(X, C)$, is as follows:

$$F(X, C) = \sum_{i=1}^{n} \sum_{j=1}^{k} m(x_i, k) p(x_i, c_k) \tag{3}$$

where $m(x_i, k)$ is an indicator function with value 1 if c_k is the closest centroid to x_i and 0 otherwise, and $p(x_i, c_k)$ is the proximity measure (distance, similarity, etc.) of x_i to c_k. Typical proximity measures used with K-means include Euclidean distance, cosine similarity, Pearson's correlation coefficient, and Kullback-Leibler (KL) divergence.[14] When $p(\cdot, \cdot)$ represents a similarity, such as cosine similarity, the function $F(X, C)$ is maximized; otherwise it is minimized. We allow the minimization (maximization) to continue until complete convergence, i.e., the algorithm completes an iteration in which the exemplars for each cluster do not change from the previous iteration.

For this paper we used a variant of K-means, called K-medoids,[15] which chooses a medoid from the data to serve as the exemplar for each cluster. The exemplar chosen is the observation for which the total pairwise distance of the points in that cluster is minimized (maximized if using similarity measures). This medoid can be viewed as the most representative example of the cluster, without the translation problems of trying to map an abstract multi-dimensional vector such as a centroid to

an appropriate example. Using medoids instead of centroids is helpful in interpreting the document similarity graphs in TopicView, as only the documents being analyzed are presented to the user.

The starting points for cluster exemplars are often chosen randomly, and to avoid local optima of $F(X, C)$ (i.e., potentially suboptimal clusterings of the data), the clustering is repeated multiple times and the best configuration selected. For our clustering, however, we choose initial exemplars using the KKZ initialization method.[16] This method selects an initial set of cluster exemplars, which often helps reduce the total number of iterations necessary for convergence of the minimization (or maximization) of $F(X, C)$ while simultaneously providing clusterings competitive with repeated random initialization and other prominent initialization strategies.[17] An additional side effect of using the KKZ initialization is that the clustering process becomes repeatable, as both the KKZ and K-medoids algorithms are deterministic.

As part of this work, all clustering experiments were performed using a variety of proximity measures: Euclidean distance, cosine similarity, Pearson correlation coefficient, KL divergence, Jensen-Shannon divergence, Chebyshev distance, Hellinger distance, Manhattan distance, and Minkowski distance.[14] The results of the clustering experiments indicated that cosine similarity and Jensen-Shannon divergence consistently led to the best clusters using the LSA and LDA models, respectively. Therefore, all images in this paper illustrating clustering results use those proximities for the two models.

In assessing which topic model leads to better performance when used as input for our clustering, we compare the true, ground-truth cluster assignments from our data with the assignments predicted by our clustering algorithm. To this end, we employed two cluster comparison measures, the RAND and Jaccard indices. Among the many difference measures for assessing and comparing clusterings, we opted for these as they are straightforward to implement and are often used as baseline measures in the literature.[18,19,20] Both RAND and Jaccard provide a measure of the similarity between two clusterings as a function of the pairwise agreement / disagreement of each document across those clusterings.

4. Data Sets

Several data sets are used throughout this paper to illustrate the various components of TopicView. These data sets are also used in the case studies demonstrating TopicView in Section 6. These data sets include synthetic data created to illustrate TopicView capabilities and differences between LSA and LDA, along with real-world data sets of news documents and newsgroup messages. All data sets have human-generated cluster labels, which are used to color-code the document groups and identify how well the algorithms' clusters match human-generated ground truth.

4.1. Alphabet *Data Set*

The ALPHABET data set consists of 26 clusters containing 10 documents each. Each cluster contains documents composed from terms starting with the same letter. The term set was constructed starting from a dictionary of 1000 words starting with the letter "a". The other letters of the alphabet were generated by prefixing each letter in turn to this base set, resulting in a vocabulary of 26,000 terms. For each of the 10 documents from each cluster, 100 terms were sampled with replacement from a uniform distribution of the corresponding 1000 terms used for that cluster. The result is a collection of document clusters that are mutually exclusive with respect to the terms used in the documents. Moreover, the terms belonging to a particular cluster are easy to identify visually, as all terms associated with a cluster start with the same letter.

4.2. DUC *Data Set*

Our second case study used the DUC data set, a collection of news articles from the Associated Press and New York Times that were used in the 2003 Document Understanding Conference (DUC) for evaluating document summarization systems.[21] The collection contains 298 documents categorized into 30 clusters, with each cluster containing roughly 10 documents focused on a particular topic or event. The categorization of the DUC data set was performed manually by a small group of subject matter experts whose task was to identify groups of documents whose main topic was similar to those in the same group. However, due to the tendancy of news articles to cover more than one topic per article, there is some unintended overlap in topics across the document groups. This characteristic of the DUC data set leads to challenges in clustering, and thus is helpful in illustrating some of the benefits of using TopicView in assessing the similarities and differences of various topic models.

4.3. Newsgroups *Data Set*

Finally, the NEWSGROUPS data set used for our third case study is a subset of the 20 Newsgroups data set which has been widely used to assess the performance of clustering and classification algorithms.[22] Our version[a] contains a subset of 10,000 email messages posted to 20 different Usenet newsgroups ranging across the main top-level subject areas of the Usenet newsgroup hierarchy. For our work, we chose a diverse subset of 6 of the original 20 newsgroups, and uniformly sampled 100 documents from the approximately 10,000 documents associated with each of those 6 newsgroups. For each message, we used all of the body text but included only the message subject from the header information, as the subject was the only header that consistently contained topical content related to the body of the message. Furthermore, we removed any message threading formatting (e.g., "`>>`") and

[a]We use a version of the 20 Newsgroups data set in which duplicate messages have been removed. This data set is available at `http://qwone.com/~jason/20Newsgroups`.

ASCII art used in user signatures posted at the ends of messages. (e.g., "=^=^=^=", "/\/\/\/\/\/", "00_oo_00", etc.). The complete set of newsgroups from the 20 Newsgroups data set are listed in Table 4.3, with the 6 newsgroups used in this paper listed in boldface.

Note that the clusters in the NEWSGROUPS data set are defined by the newsgroup names, but the messages posted to a particular newsgroup may not be completely or even partially related to the subject matter of that newsgroup. Because there was no moderation of these groups during the time period of the posting data, there is no guarantee that the individual messages conform to the Usenet newsgroup hierarchy in any way. For this reason, as with the DUC data, this data set poses challenges for clustering algorithms when using the newsgroup as the true cluster assignment. Thus, the NEWSGROUPS data set is useful in illustrating the different components of TopicView and how they help in understanding relationships between documents in text collections.

Table 1. Names of Usenet newsgroups used in the 20 Newsgroups data set.

comp.graphics	sci.electronics
comp.os.ms-windows.misc	**sci.med**
comp.sys.ibm.pc.hardware	sci.space
comp.sys.mac.hardware	**misc.forsale**
comp.windows.x	talk.politics.misc
rec.autos	**talk.politics.guns**
rec.motorcycles	talk.politics.mideast
rec.sport.baseball	talk.religion.misc
rec.sport.hockey	alt.atheism
sci.crypt	**soc.religion.christian**

Note: The newsgroups listed in boldface are those used in the work presented in this paper.

5. TopicView

TopicView is a standalone application enabling comparisons of multiple topic models and their impact on the task of document clustering via visual inspection, exploration, and analysis in several components and views. The graphical user interface (GUI) for TopicView was designed to facilitate extensibility (in terms of adding additional model or analysis views) and coordinated exploration (across the different types of analyses available). In this section, we present an overview of TopicView, followed by specific details of the various components currently available. See Figure 1 for an overview of the system.

At the start of a new analysis session, users load a text corpus into TopicView, which performs several pre-processing steps on the data prior to input to the LSA and LDA model creation pipelines (see Section 3 for more details about these models). Identical LSA concept and LDA topic counts are selected by the user to gen-

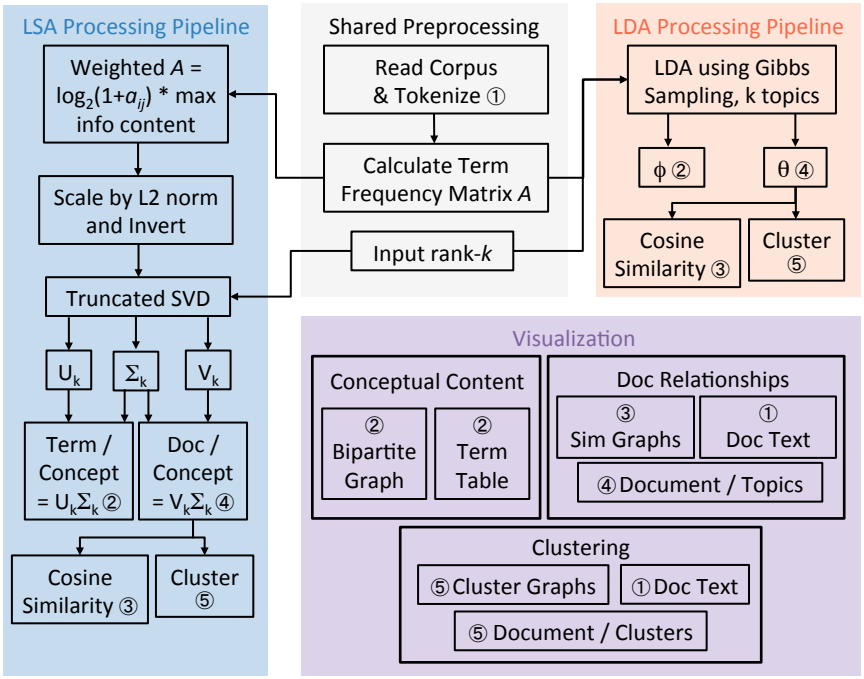


Fig. 1. TopicView system diagram. Numbered outputs in the LSA, shared preprocessing, and LDA pipelines connect to numbered inputs in the visualization diagram.

erate matching numbers of concepts and topics for their respective models. The document relationships computed using these two models use the same cosine similarity, edge threshold, and graph layout components to produce document similarity graphs. Our goal throughout this process is to limit differences to just those attributable to the different modeling approaches, independent of the pre-processing and transformations of the data required to move from raw text data to data structures that can be used as input to the model fitting and graph layout algorithms.

Conceptual content, document relationships, and clustering results are visualized using separate panels. The panels are designed to enable exploration of progressively more detailed relationships from the corpus level down to individual document text. For consistency, LSA-generated components are always displayed in blue and appear on the left (left nodes in the bipartite graph, left-most columns in tables, and left document similarity graphs), whereas LDA-generated model components are displayed in red and appear on the right. The choice of colors and placement was arbitrary and could be easily changed, e.g. if a different topic model were used in the analysis instead of LSA or LDA.

In the next few sections, the main panels are described in greater detail.

5.1. *Conceptual Content Panel*

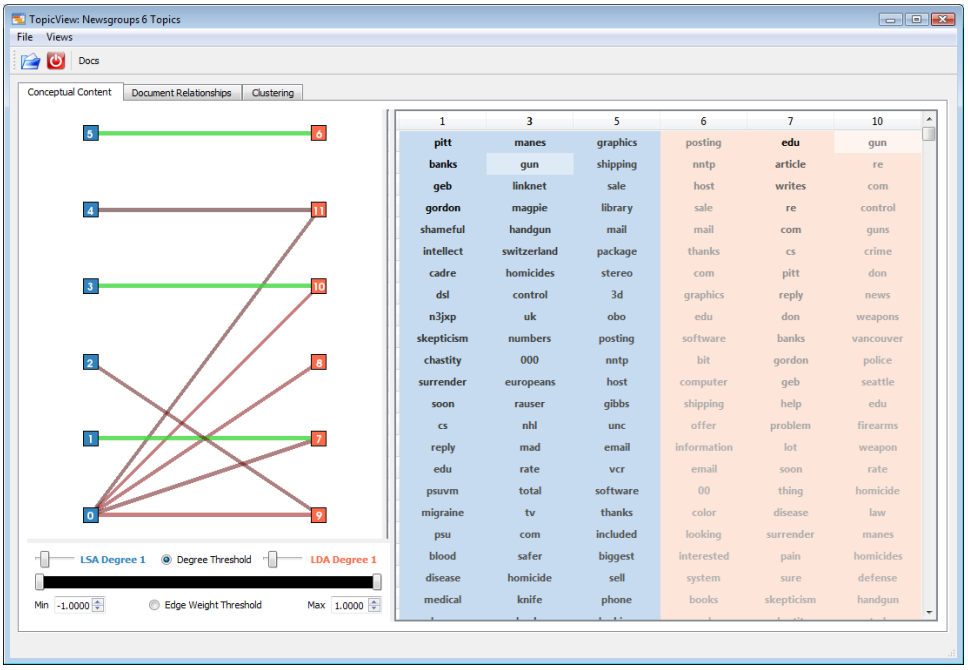


Fig. 2. TopicView's Conceptual Content panel containing *Bipartite Graph* (left) and *Term Table* (right) views.

At the highest level, we want to know how concepts (from LSA) compare to topics (from LDA), while hiding any details of the underlying models or methods. A bipartite graph provides an abstract overview of these relationships by connecting concepts and topics with weighted edges. Conceptual content is represented by the relative importance of the terms within each concept/topic. Although we cannot directly compare the weightings assigned to individual terms between concepts and topics (i.e., correlations from LSA and probabilities from LDA), we can visually compare the ordering and relative weighting of the terms.

5.1.1. *Bipartite Graph*

Ideally, if there were a one-to-one relationship between concepts and topics, we would want a representation that made the correspondence explicit via visual pairing. On the left side of the *Conceptual Content* panel in Figure 2, the *Bipartite Graph* provides this pairing by horizontally aligning strongly correlated pairs of concepts and topics and connecting them with a line that is color-coded based on the strength of the correlation).

Concept/topic similarities are calculated as follows:

(i) Scale LSA's left singular matrix by its singular values ($V_k\Sigma_k$).
(ii) Concatenate the result with LDA's ϕ matrix.
(iii) Compute the matrix of pairwise cosine similarities of all rows of the concatenated matrix.
(iv) Truncate the resulting similarity matrix to be just the upper right quadrant, retaining only similarities between unique pairs of LSA concepts and LDA topics.
(v) Sort the truncated edge list in descending order.

The edge weights are color-coded from blue (-1) to black (0) to red (1) to preserve the distinction between positive and negative similarities.

Fixing the node positions for the left-hand (LSA) model in their rank order, we use a greedy approach for placing nodes on the right-hand (LDA) model relative to the fixed nodes on the left. Nodes are laid out in declining edge weight order so as to draw the strongest pairwise similarities between concepts/topics in the different models with horizontal edges.

We provide two interactive filtering mechanisms for reducing the number of *Bipartite Graph* edges shown, *Degree Threshold* and *Edge Weight Threshold*, the controls for which are visible below the graph. *Degree Threshold* independently controls the minimum vertex degree for each side of the bipartite graph with a separate slider. Edges are chosen in descending weight order so that strongest edges are always visible first (i.e., for degree 2, the two strongest edges are drawn). Although the sliders control the minimum number of visible edges for each node, some nodes (such as LSA node 0) may exceed the minimum due to edges that exceed an adjacent node's minimum edge count. Alternatively, *Edge Weight Threshold* displays all of the edges whose weights fall within a user specified range.

Nodes and edges in the *Bipartite Graph* view can be selected, with selections shown in green. Selecting an edge selects its two connected nodes. Nodes correspond to columns in the *Term Table* and columns in the *Document Table* on the *Document Relationships* panel. As shown in Figure 2, selections limit the columns displayed in both tables to just the selected concepts and topics. Then, column adjacency can be used to compare word lists in the *Term Table* and see which documents contribute most heavily to those concepts/topics in the *Document Table*. Clicking anywhere outside the graph clears the current selection and makes every column visible.

5.1.2. *Term Table*

The terms associated with each concept/topic, listed in decreasing order of importance, are presented in the *Term Table* on the right side of the *Conceptual Content* panel in Figure 2. Text color provides an additional cue about the relative weights of terms, varying from black for the highest weights to light gray for the lowest. Since we are most interested in distinguishing weighting differences at the high end of the scale, we use a logarithmic mapping that increases the number of luminance steps as we approach black. Since the LSA and LDA ranges are independently scaled

based on their values, luminance differences cannot be directly compared between models.

Individual terms within the table are selectable. Once selected, each instance of that term within every concept/topic is highlighted with a lighter background. The selection is linked to the *Document Text* view, so that every instance of the term within the selected documents is displayed in red. This provides users with the ability to quickly determine the distribution of a single term across concepts or topics both within and across the different models.

5.2. *Document Relationships Panel*

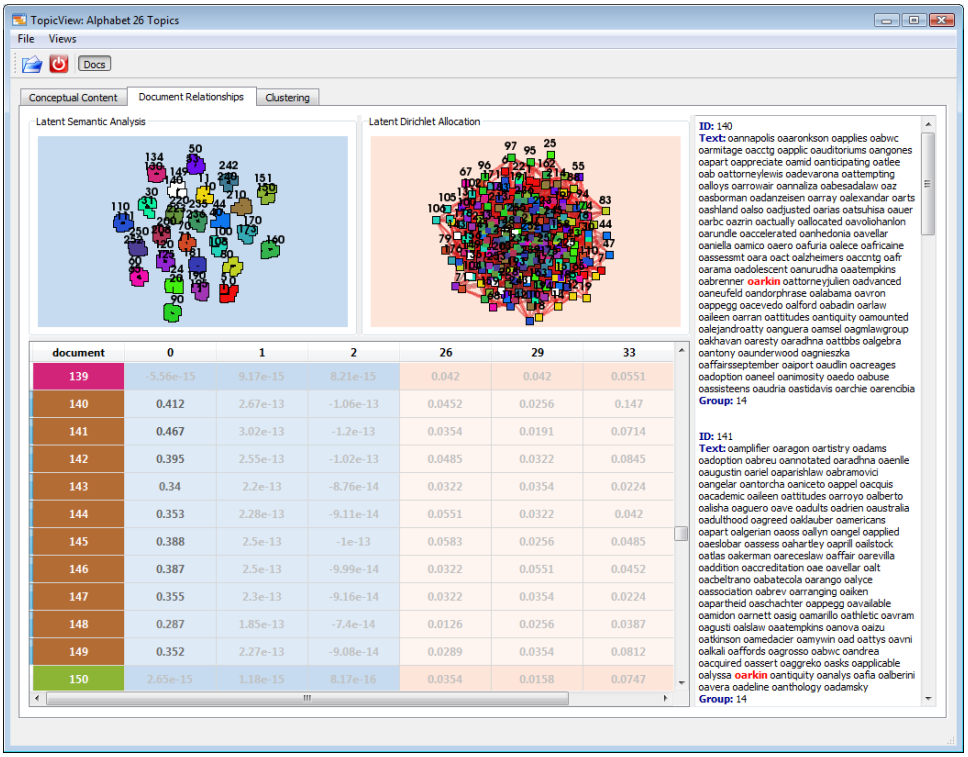


Fig. 3. TopicView's Document Relationship panel containing *Document Similarity Graphs* (top left), *Document Table* (bottom left) and *Document Text* (right) views.

Document groupings as shown in *Document Similarity Graphs* provide an alternative view of LSA and LDA model differences. Because similarity graphs are often used as a proxy for true clustering, we are interested in those cases where a set of documents exhibits strong links between members of the group and weak links outside the group. Although there is a tendency to try to identify concepts/topics with these groups, the weightings shown in the *Document Table* demonstrate that doc-

ument groups frequently contribute in varying degrees to multiple concepts/topics (weightings spread across rows). Similarly, concepts/topics typically include multiple document groups (weightings spread across columns). The visual combination of the graphs and tables on the same panel enables the user to locate and select the documents associated with either conceptual content or groups, then read their full text in the *Document Text* view.

5.2.1. *Document Similarity Graphs*

We compute cosine similarities for LSA using the right singular vectors, scaled by the singular values. For LDA, we compute cosine similarities using the θ matrix. This generates edges between every pair of documents, so we reduce visual clutter by thresholding edges based on their weight. We want to keep the strongest links, while at the same time providing some connectivity to all documents. We determine which edges to keep on a document-by-document basis as follows:

(i) Sort the set of edges associated with each document node in descending order by weight.
(ii) Keep all edges with weights greater than a significance threshold (we use 0.9).
(iii) If the number of highly weighted edges for a document is less than a specified count (5 in all of our examples) continue adding edges in diminishing weight order until that count is reached.

We use a linear time force-directed layout for the graphs. As shown in the upper left corner of Figure 3, each document is labeled with its ID and color-coded using a ground-truth category. Edge color saturation indicates similarity weight, with low values in gray and high values in red. Nodes and edges can be selected, with node selections drawn in white and edge selections drawn in blue. The LSA and LDA graphs are linked, so corresponding selections are shown in both (note that some edges may exist in one graph and not the other). Additionally, the LSA and LDA *Document Cluster Graphs* are also linked, so corresponding selections are highlighted within the context of each algorithm's clusters. The selected documents are also highlighted in the *Document Table* and *Cluster Table*, while their full text is displayed in the *Document Text* view.

5.2.2. *Document Table*

The *Document Table* (lower left Figure 3) shows the concatenation of the transpose of LSA's right singular vector, scaled by its singular values, and LDA's θ matrix. In a manner identical to the *Term Table*, the values in the table are varied between black and light gray to permit rapid visual scanning of rows and columns to find darker, more highly weighted documents. This facilitates comparisons of the relative significance of documents within a set of concepts or topics. Selecting rows within the table will highlight nodes in all graphs and display the selected document contents

in the *Document Text* views.

5.2.3. *Document Text*

The *Document Text* view provides the full text contents of multiple documents, selected using the *Document Similarity Graphs*, *Document Cluster Graphs*, *Document Table*, or *Cluster Table*. Each document is displayed as three fields: the document ID, the raw document text, and a ground-truth categorical ID. If a term is selected in the *Term Table*, that term is highlighted in red throughout the raw text. When displaying longer documents and multiple documents, the view can be scrolled.

5.3. *Clustering Panel*

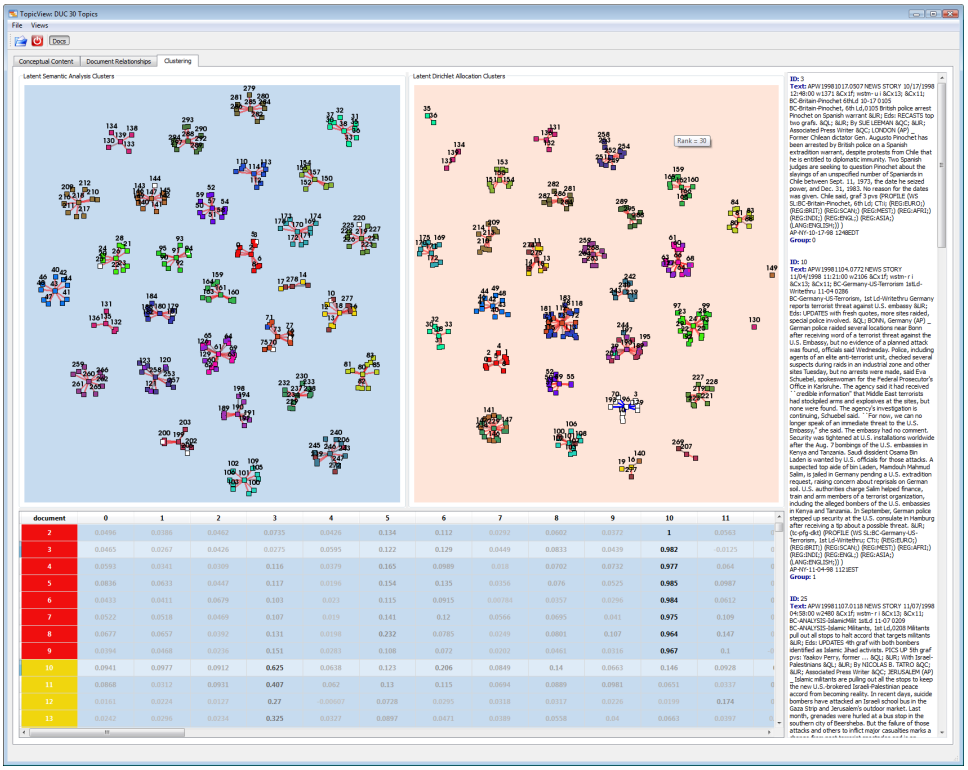


Fig. 4. TopicView's *Clustering Panel* containing *Document Cluster Graphs* (top left), *Cluster Table* (bottom left) and *Document Text* (right) views.

While the first two user interface tabs provide visualizations that directly compare the models generated by LSA and LDA, the third tab explores how model differences impact downstream algorithms that use the generated models as inputs. In particular, we examine clustering and how differences in the models effect cluster

results. The clustering algorithms identify clusters and their associated medoids, generating a matrix of relationships between each document and each medoid, and the relationships of the medoids to each other. These can be visualized as a graph that uses both intra- and inter-cluster proximities to provide a high-level view of the corpus content, i.e. what clusters exist and how closely related they are to one another. The visualization includes three views, the *Document Cluster Graphs*, a *Cluster Table*, and *Document Text*. The *Document Text* is identical to that included in the *Document Relationships* panel.

5.3.1. *Document Cluster Graphs*

Although the clustering output provides relationships between each document and each medoid, visualizing the full graph provides little insight due to the dense overlap of the edges. We reduce visual clutter by rendering only the edges between each document and the medoid with the greatest proximity, combined with only those edges between medoids whose proximity exceeds a significance threshold (we use 0.9). We use the same linear force-directed layout as we did with the *Document Similarity Graphs*, and the same edge and node color-coding and selection interfaces.

5.3.2. *Cluster Table*

The *Cluster Table* shows the concatenation of LSA and LDA's document-cluster proximity matrices. For each cluster, the proximities between each document and the medoid representing the cluster are displayed. The proximity values determine the darkness of each number's rendering, as was done in both the *Term Table* and *Document Table*. Selections in this table are linked to all four graphs, the *Document Table*, and both *Document Text* views.

6. Case Studies

In this section, we present the results of using TopicView's visual analysis and comparison capabilities to illustrate similarities and differences between two topic models: LSA and LDA. We use both synthetic and real-world corpora to demonstrate the various characteristics of the visual components and how they can be used in model assessment and comparison. The goals of the case studies include:

- Illustrating the use of TopicView's coordinated, multi-view capabilities for efficient navigation of relationships between LSA and LDA models.
- Determining the relationship between LSA and LDA models with respect to the most important terms and overall term distributions associated with topically-related groups of documents.
- Identifying strengths of the LSA and LDA models with respect to document clustering and model interpretability.
- Illustrating how visual analysis capabilities can aid in the understanding of relationships in large document collections.

For all results presented here, the topic models (LSA and LDA) were computed using the ParaText library[23] from the open source Titan Informatics Toolkit.[24] For LDA, we set $\alpha = 50 \ / \ K$, $\beta = 0.1$, and used 1 sampling and 200 burn in iterations. For both LSA and LDA, the number of concepts and topics, respectively, are chosen by the user; results presented below explicitly identify the values used for this parameter, which is always shared by the two models to reduce any differences due to the number of dimensions used in the models generated. Clustering results were also computed using Titan.

6.1. *Synthetic Data Studies*

In this section, we report the results of applying the visual analysis capabilities in TopicView to the ALPHABET data set (see Section 4.1), a collection of synthetic documents with independent term distributions across documents in 26 clusters.

As LSA concepts are, by definition, orthogonal latent feature vectors, we expect that a 26-concept LSA model should be able to identify each of the document clusters in the ALPHABET data set using a single concept per letter (i.e., one latent feature for each of the sets of terms beginning with the same letter). Note that for many real-world document collections the optimal number of clusters is not known *a priori* and documents related to a particular topic do not consist of terms unique to that topic alone. However, the purpose of this study is to demonstrate the use of visual analysis to identify differences in the LSA and LDA models when one model (i.e., LSA) is able to exactly cluster the data.

Figure 5 presents the *Document Similarity Graphs* for a 26-concept LSA model (left) and 26-topic LDA model (right) of the ALPHABET data set. As predicted, we see that the LSA model consists of independent concepts that lead to a perfect clustering of the documents; there are 26 disconnected components in the graph, and each component consists of nodes colored with the same ground-truth cluster label. On the other hand, LDA has difficulty modeling the term relationships across documents, generating strong relationships across the entire collection. As LDA models are mixture models and there is absolutely no term mixing across documents in different clusters in the ALPHABET data set, LDA is not a suitable model for such data. Such differences may not be so obvious in real-word document collections, but the amount of mixing (i.e., diversity in distribution) of terms across documents and/or topically related groups of documents may significantly impact the suitability of a particular topic model.

As shown in Figure 6, the *Conceptual Content* panel can be used to better understand the differences between term distributions within the LSA concepts and LDA topics. Through selections in the *Bipartite Graph*, the concept/topic columns have been limited to the three LSA concepts associated with the largest singular values (i.e., concepts 0, 1, and 2) and the most related LDA topics in terms of cosine similarity (topics 33, 29, and 26, respectively). In the *Term Table*, we can see that words beginning with "o", "v", and "e" are most highly correlated with LSA

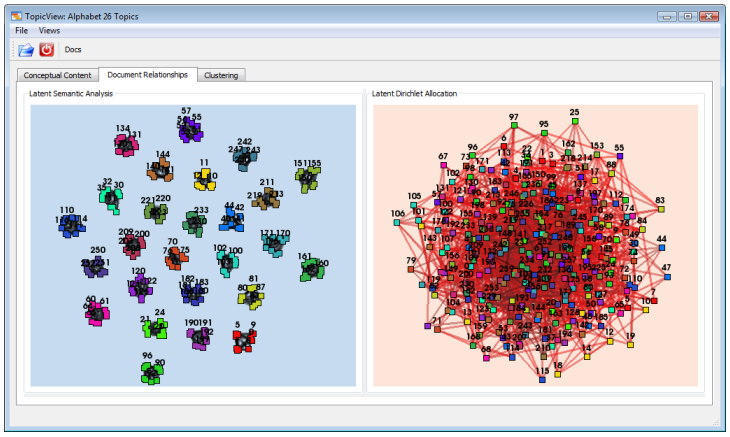


Fig. 5. *Document Similarity Graph* views depicting document relationships modeled using LSA (left) and LDA (right) for the ALPHABET data set.

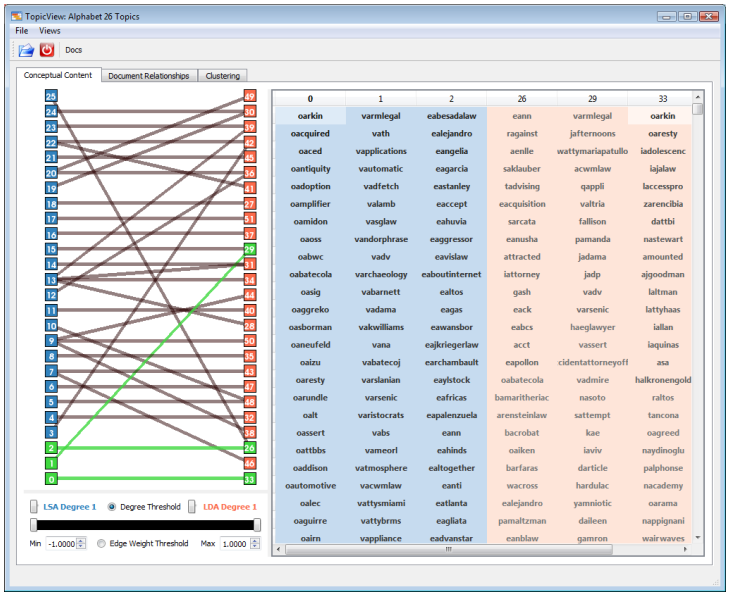


Fig. 6. Conceptual Content panel views depicting relationships between LSA concepts (blue) and LDA topics (red) for the ALPHABET data set. The LSA concepts identify the independent term sets in the data, whereas terms starting with different letters are highly mixed across LDA topics.

concepts 0, 1, and 2, respectively. In contrast, the terms with highest probability of being part of LDA topics 33, 29, and 26 contain some terms beginning with those letters, but there is no clear connection to any particular document cluster. Further investigation using the *Document Table* and *Document Text* views in the *Document Relationship* panel confirm that LSA models the clusters correctly; see Figure 7 for

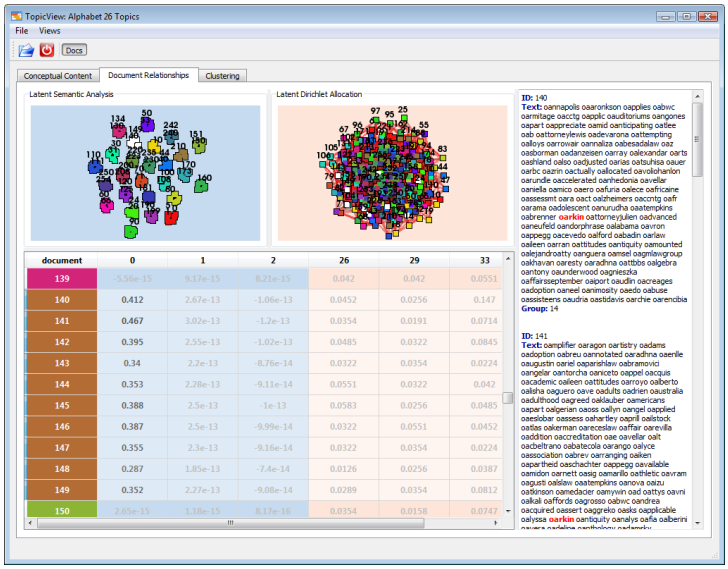


Fig. 7. For the ALPHABET data set, the *Document Table* and *Document Text* views in the *Document Relationships* panel show that LSA concept 0 contains only "o" documents (140-149), whereas LDA topic 33 has a mixed set of weights for that cluster.

an illustration of how these views are used to verify that only "o" documents are related to LSA concept 0.

Once it was established that LSA was modeling the clusters accurately and LDA was not, we used the *Bipartite Graph* to identify relationships between LSA concepts and LDA topics. Figure 8 shows the *Bipartite Graph* depicting relationships between LSA (blue nodes) and LDA (red nodes) models in terms of cosine similarity between the concept vectors (i.e., left singular vectors) and topic vectors (i.e., rows of ϕ), respectively. The left and right images in the figure show the graph edges thresholded by degree and edge weight, respectively. In both images, we see that all of the relationships between LSA concepts and LDA topics are weak as indicated by the gray edges (recall that red edges would indicate strong relationships). The left image depicts the strongest connections for LSA concepts (i.e., LSA degree threshold of 1 and LDA degree threshold of 0). In this image, we can quickly see that there are some LDA topics (e.g., 28, 41, 44, 46, and 49) that are unrelated to any LSA concepts, each of which models one of the 26 distinct document clusters. Furthermore, we see that at an edge-weight threshold of 0.1451 (i.e., the maximum threshold value for which all LDA topics are related to at least one LSA topic), most of the LDA topics are more strongly connected to several LSA concepts (i.e., relatively high out degree on LDA topic nodes) before any one LDA topic is related to LSA topic 10. This is a further indicator that LDA is not modeling the term relationships accurately within each document cluster.

These outcomes are consistent with our expectations for this synthetic data set.

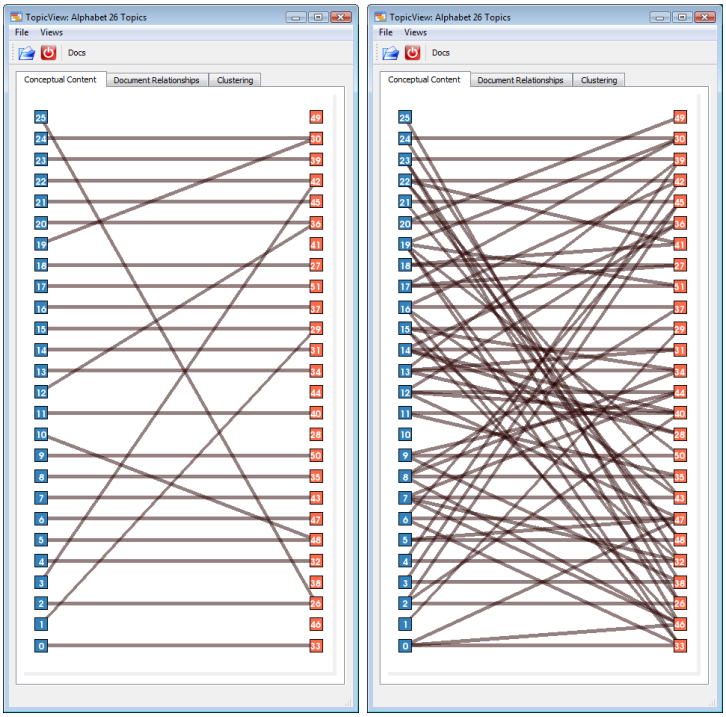


Fig. 8. *Bipartite Graph* views depicting conceptual content relationships for the ALPHABET data set modeled with LSA (blue nodes) and LDA (red nodes) using a minimum degree threshold of 1 for LSA concept nodes (left) and a minimum edge threshold of 0.1451 (right).

Because documents from different clusters are entirely disjoint, each cluster is well approximated by a unique singular vector from the LSA model, leading to high correlation between documents within a cluster and very low correlation across clusters. Conversely, this disjunction represents a very difficult case for LDA implementations using collapsed Gibbs sampling. At an intuitive level, such LDA model fitting methods rely on co-occurrence of terms between documents to guide a random walk toward more probable topic configurations. In the ALPHABET data set, explicit term co-occurrence between clusters has been suppressed completely. Moreover, the small document size (relative to dictionary size) and uniform sampling strategy results in a low degree of overlap between documents within a cluster. In such a situation, the topics from LDA tend to be random with more or less uniform term distributions. In further work, we plan to explore these characteristics in greater detail, assessing whether the diversity of term distributions across documents (and, more specifically, across groups of topically related documents) could lead to indicators of model applicability to particular document collections.

6.2. *Real-World Data Studies*

As discussed in the previous section, the ALPHABET data set was useful in illustrating the differences across the LSA and LDA models. However, the structure and relationships among those synthetic documents is not representative of more typical document collections. Even when corpora contain clusters of documents across a wide range of disparate subject areas, there is often significant overlap in vocabulary across documents in different clusters.

In this section, we present the results of case studies involving two real-world data sets: the DUC (Section 4.2) and NEWSGROUPS (Section 4.2) data sets. Recall that the DUC data set has more document clusters (30 versus 6) with fewer documents in each cluster (about 10 versus 100) and a different number of unique terms (13,641 versus 15,723) than the NEWSGROUPS data set. Thus, these two data set sets can be used to identify how the visual analytics presented in this paper perform when applied to document collections of different sizes with different patterns of term distributions within documents and across document clusters.

6.2.1. *Topic Model Comparisons: How Can Visual Analysis Help?*

To investigate the relationships between LSA and LDA modeling on a real-world collection of documents, we applied TopicView to the DUC data set. The DUC data set was assembled for a summarization contest, so it contains subsets of documents whose general topics appeared very different to the annotators. Using TopicView, we demonstrate how LSA and LDA models are similar in identifying clusters with consistent term distributions across documents in particular clusters, but differ in modeling weak connections between document clusters.

We begin our analysis of the relationships between the LSA and LDA models by examining the document similarity graphs for the two models. Figure 9 shows the *Document Similarity Graphs* view for a 30-concept LSA model (left) and 30-topic LDA model (right). The LSA model results in a graph with 14 disconnected components, 13 of which exactly match clusters identified by the human annotators. The large connected component located in the center of the layout indicates that have some degree of toverlap in their term distributions. Note that there are many highly interconnected, compact subgraphs which correspond to true document clusters (as indicated by the node colorings) that are connected to other such subgraphs by one or two edges. Even without the node colors, the graph topology indicates that there are highly related documents in these subgraphs and we can trace the connections between the subgraphs through specific documents by edges connecting them. Thus, we conclude that the LSA model provides a useful clustering of the documents in the DUC data set.

When using the shared layout to view relationships between documents as computed using the LDA model (as shown on the top right in Figure 9), we see that there are many inter-cluster relationships identified. This indicates that LSA and LDA are clearly modeling different characteristics of the data. When the LDA-specific

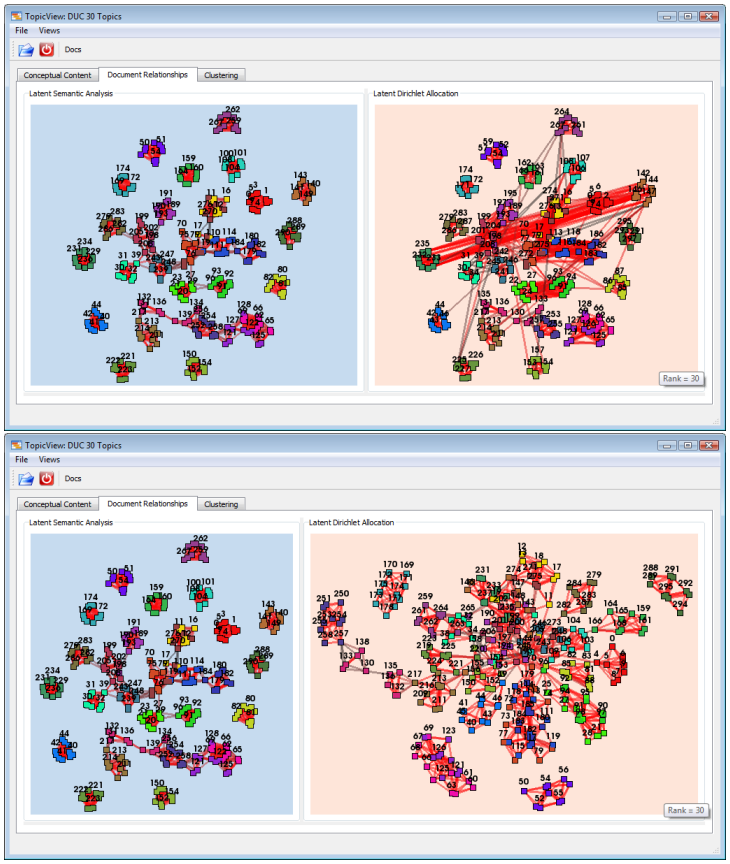


Fig. 9. *Document Similarity Graph* views depicting document relationships modeled using LSA and LDA for the DUC data set, using a shared node layout in the upper image and non-shared layouts in the bottom image.

layout is viewed (as shown on the bottom right in Figure 9), we get a much better sense of the clustering produced by the LDA model; there are disconnected components (indicating document clusters) and a subgraph structure similar to the LSA model. Thus, we conclude that the LDA model also provides a useful clustering. However, other visual analytics provided in TopicView can be used in conjunction with these *Document Similarity Graphs* to explore in more detail the similarities and differences between the tightly connected subgraphs.

In Figure 10, TopicView's *Bipartite Graph* shows the relationships between LSA concepts and LDA topics. Using the degree (left) and edge weight (right) thresholding controls, we see there is a strong, unique relationship between LSA concept 0 and all LDA topics relative to the other pairwise concept/topics relationships. This relationship is due to the fact that LSA concepts (i.e., left singular vectors, U_k from Eq. (1) in the model description in Section 3.1) account for decreasing amounts of

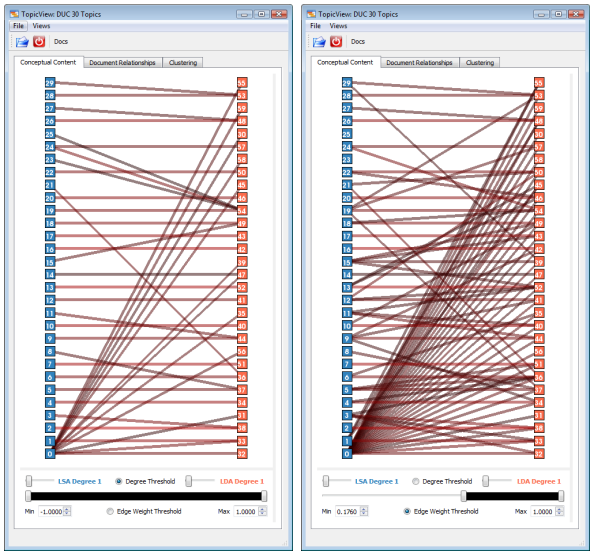


Fig. 10. *Bipartite Graph* views depicting conceptual content relationships for the DUC data set modeled with LSA (blue nodes) and LDA (red nodes) using a minimum degree threshold of 1 for LSA/LDA nodes (left) and a minimum edge threshold of 0.1760 (right).

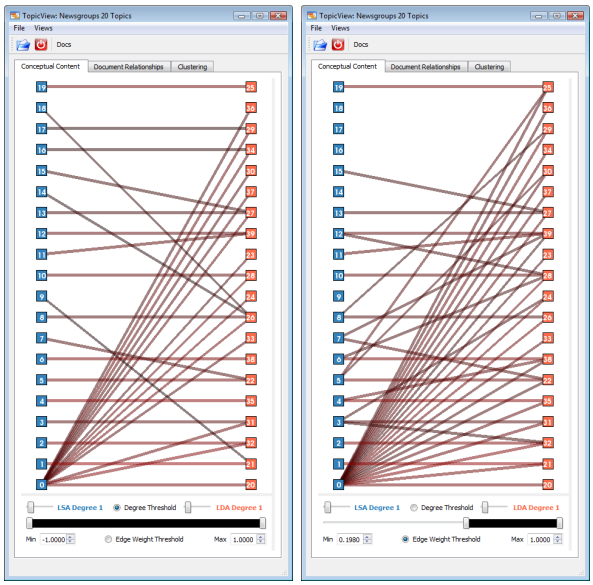


Fig. 11. *Bipartite Graph* views depicting conceptual content relationships for the NEWSGROUPS data set modeled with LSA (blue nodes) and LDA (red nodes) using a minimum degree threshold of 1 for LSA/LDA nodes (left) and a minimum edge threshold of 0.1980 (right).

variance of terms across the documents. Thus, LSA concept 0 acts as a generic concept that summarizes all of the main term-term and term-document interactions.[1] Moreover, we see this same behavior with LSA concept 0 across a range of dimension choices (values of k used to generate the models) in each of the data sets we have studied (see Figure 11 for an example of the NEWSGROUPS data set with 20 concepts/topics), providing evidence of this being a general characteristic of the LSA models.

6.2.2. *Model Accuracy: Is a Single Model Adequate?*

Data analysts using our analysis capabilities often ask us whether one topic model better expresses the relationships in the data over another. Determining the answer to that question is a challenging problem, as it often depends on the types of questions being asked of the data and the specific types of modeling and analysis that will be performed using the topic models. Using TopicView, we find evidence that the answer may in fact be that we want to use both (or all) topic models available for a given analysis task, as the different models may be able to identify some relationships in the data that the other(s) may not, and vice versa. In this section, we present such evidence for the DUC data set.

As we see in the *Bipartite Graph* on the left in Figure 10, there are several LSA concepts strongly connected to a single LDA topic or vice versa. Two such examples include (a) LSA concepts 9 and 11, which are strongly connected to LDA topic 44, and (b) LSA concepts 6 and 21, which are strongly connected to LDA topic 36. Figure 12 shows the top 10 terms (in descending order by the coefficients and/or probabilities associated with those terms in the topic models) in the LSA concepts and LDA topics for these two cases (case (a) on the left and case (b) on the right).

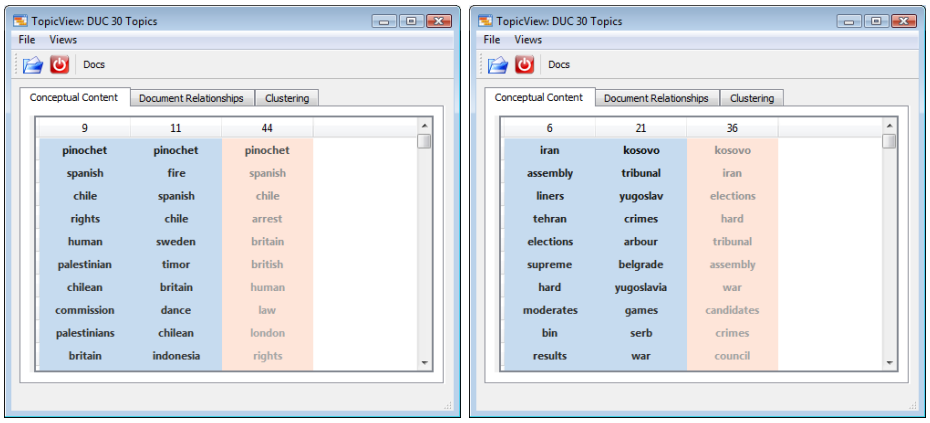


Fig. 12. Top 10 terms associated with the concepts/topics for the DUC data set, where multiple LSA concepts are strongly connected to a single LDA topic.

In case (a), LSA concept 9 along with LDA topic 44 appear related to the cluster of documents about the Chilean leader Pinochet, whereas LSA concept 11 has combined the Pinochet cluster with clusters of documents about a dance hall fire in Sweden and political unrest in Timor, which do not appear related. Using the *Document Table* and *Document Text* views, though, we find that the full LSA concept vectors for concepts 9 and 11 are negatively correlated for all documents except those in the Pinochet cluster and two other sets of documents, one related to the political unrest in Timor (concept 13) and one related to war crimes by Serbian leadership (concept 21). Tracing the terms used in those documents, we find that there are documents in the DUC data set containing terms that span these apparently different concepts, thereby accounting for the connections between Pinochet and the Swedish fire and unrest in Timor. For example, document 87 contains the terms "Pinochet,", "Chile," "Timor," "Indonesia," and "Britain"; and document 121 contains the terms "Spanish," "fire," "chile," and "Britain". As discussed in Blei et al.,[2] LDA handles polysemous term usage much better than LSA (Chile the country versus fire roasted chile versus a fire in Sweden). We conclude that LSA is modeling the Pinochet cluster well with concept 9, as well as the more subtle, polysemous cross-cluster term relationships between Pinochet, the Swedish fire and Timor politics with concept 11.

Further inspection of LSA concept 21, which was identified above as being related to concept 11, shows that it is also involved in the multiple LSA concept, single LDA topic relationships in case (b). Although LSA concepts 6 and 21 model the two clusters of documents about elections in Iran and war crimes in Serbia well, these appear to be combined in LDA topic 36. Following the same exploration performed for case (a), we find that there are many documents in different clusters regarding politics in different areas of the world. These documents can be found by either exploring the *Document Similarity Graphs* or by combined use of the *Document Table* and *Document Text* views to identify term relationships leading to the combined LDA topic.

We conclude from these cases that LSA and LDA model the most tightly coupled document clusters well (as indicated by many strong horizontal edges in Figure 10), but model more subtle relationships between documents (and thus clusters) in different ways. We conjecture that many of the differences are due to the weighting values used in the two models. LSA's singular vectors involve both positive and negative weights for a term (as they represent correlations). This allows components of one concept to "cancel out" components of another, thus allowing linear combinations of two or more concept vectors to account for separate clusters that are related in their use of common terms. By comparison, the term weights in LDA topics are strictly positive (as they represent probabilities): there is no way to decrease a term's probability by adding in parts of other topics. Both of these approaches have their individual strengths. Using the visual analytics presented in this work, we were able to quickly identify and explore these differences. Moreover, using the

multiple linked views providing diverse perspectives of the models and relationships in the document collections, we were able to specifically identify the documents and terms that led to the model differences.

These differences do not lead to a clear answer to the original question posed at the start of this section. In fact, the results presented here indicate that a combination of the relationships identified in data using different topic models may lead to better understanding of document collections than using the models individually. This combined perspective has shown great promise in the area of data categorization and heterogeneous ensemble modeling, where multiple types of classification models are combined in ways to outperform the individual models. Although similar challenges to ensemble model combination in that area would need to be addressed when attempting to combine topics models for improved performance, this line of investigation may lead to new challenges that are quite specific to the latter. In future work, we plan to investigate this question of whether topic models can be combined in ways to lead to improved analysis in tasks such as document clustering, classification, and summarization.

6.2.3. *Short Documents: Why Do They Matter So Much?*

In the final real-world case study presented here, we illustrate a behavior of LDA models that can have a significant impact on the interpretability of the model in terms of the *Document Similarity Graphs* and effects on clustering. This behavior was originally explored and discussed by the authors for the DUC data set,[25] and that discussion is expanded here, including corresponding analysis of the NEWSGROUPS data set, which exhibits similar behavior. Specifically, we find that very short documents impact the ability of LDA to model some topically related content accurately.

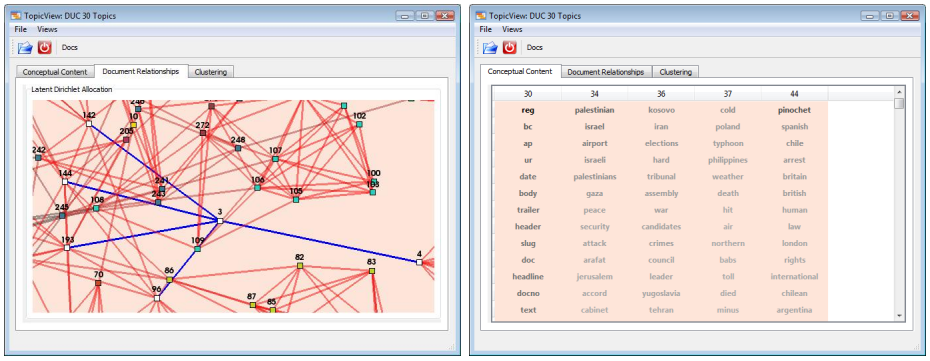


Fig. 13. Left: The *Document Similarity Graph* view displays document 3 and the connections to its immediate neighbors highlighted in white (nodes) and blue (edges). Right: The *Term Table* view displays topics associated with the selected documents shown in the *Document-Topic Table* images in Figure 14. The topics are chosen based on the strength of the document-topic weights for the selected documents.

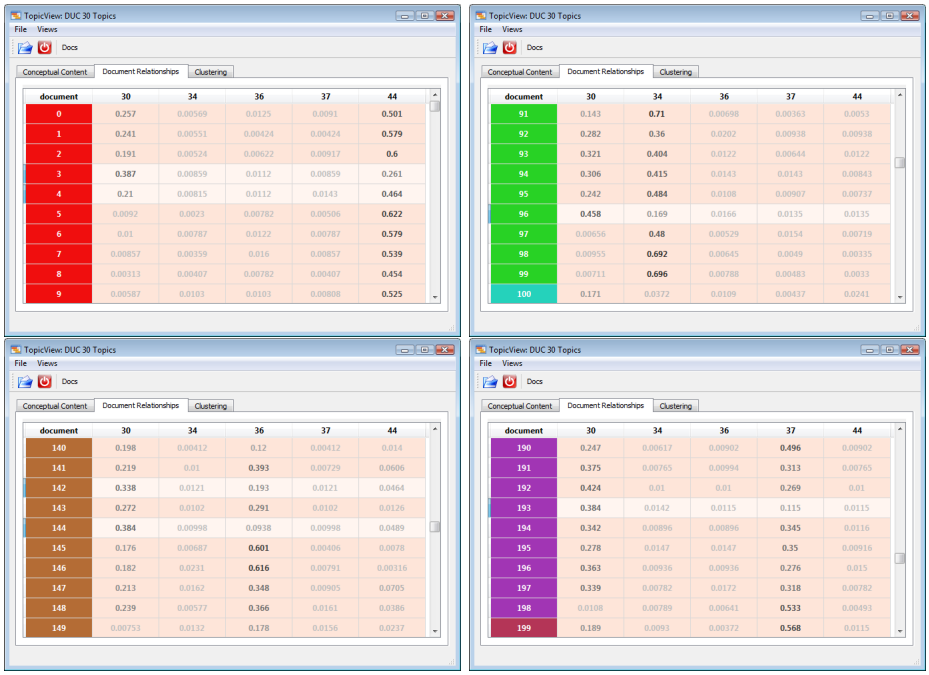
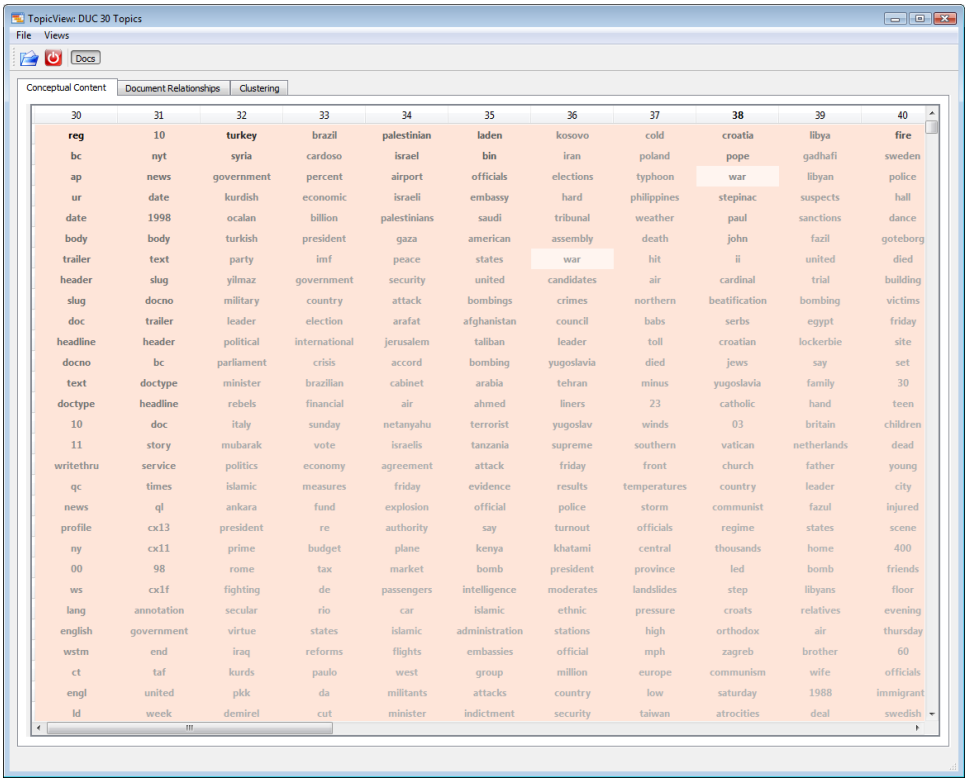


Fig. 14. The document-topic weights for document 3 (top left) and its selected neighbors.

In working with the LDA *Document Similarity Graphs* for the DUC data, we were struck by the high degree of connectivity in the graph. Many of the edges were unexpected, connecting documents covering seemingly unrelated topics. Document 3 shown in the center of the graph in Figure 13 provides a good example of this. According to the text of documents 3 and 4, both are stories about Pinochet's arrest in Britain. Document 96 is a story about Israel delaying flights from a Palestinian airport. Documents 142 and 144 are about a Yugoslavian tribunal to prosecute Bosnian war crimes. Finally, document 193 is about cold weather killing 39 people in Moscow. Why is LDA linking these documents? We felt that understanding the cause of these links was essential, since their presence alters both the graph layout and its interpretation by users.[26]

The stories in the selected documents are represented by topics 44, 34, 36, and 37, respectively. Terms capturing these key concepts can be seen amongst the top terms for these topics in Figure 13 (right). The selection of these four topics comes from an examination of the document-topic weights for each document shown in Figure 14. We have included all topic columns that have darker/higher weights within the selected rows, including topic 30, whose most significant terms do not match any of the story lines seen in the documents' texts. Examining the terms in topic 30, we find that XML document tags ("headline", "slug", "body", etc.) predominate. As shown in Figure 15, topics 30 and 31 capture Associated Press



30	31	32	33	34	35	36	37	38	39	40
reg	10	turkey	brazil	palestinian	laden	kosovo	cold	croatia	libya	fire
bc	nyt	syria	cardoso	israel	bin	iran	poland	pope	gadhafi	sweden
ap	news	government	percent	airport	officials	elections	typhoon	war	libyan	police
ur	date	kurdish	economic	israeli	embassy	hard	philippines	stepinac	suspects	hall
date	1998	ocalan	billion	palestinians	saudi	tribunal	weather	paul	sanctions	dance
body	body	turkish	president	gaza	american	assembly	death	john	fazil	goteborg
trailer	text	party	imf	peace	states	war	hit	ii	united	died
header	slug	yilmaz	government	security	united	candidates	air	cardinal	trial	building
slug	docno	military	country	attack	bombings	crimes	northern	beatification	bombing	victims
doc	trailer	leader	election	arafat	afghanistan	council	babs	serbs	egypt	friday
headline	header	political	international	jerusalem	taliban	leader	toll	croatian	lockerbie	site
docno	bc	parliament	crisis	accord	bombing	yugoslavia	died	jews	say	set
text	doctype	minister	brazilian	cabinet	arabia	tehran	minus	yugoslavia	family	30
doctype	headline	rebels	financial	air	ahmed	liners	23	catholic	hand	teen
10	doc	italy	sunday	netanyahu	terrorist	yugoslav	winds	03	britain	children
11	story	mubarak	vote	israelis	tanzania	supreme	southern	vatican	netherlands	dead
writethru	service	politics	economy	agreement	attack	friday	front	church	father	young
qc	times	islamic	measures	friday	evidence	results	temperatures	country	leader	city
news	ql	ankara	fund	explosion	official	police	storm	communist	fazul	injured
profile	cx13	president	re	authority	say	turnout	officials	regime	states	scene
ny	cx11	prime	budget	plane	kenya	khatami	central	thousands	home	400
00	98	rome	tax	market	bomb	president	province	led	bomb	friends
ws	cx1f	fighting	de	passengers	intelligence	moderates	landslides	step	libyans	floor
lang	annotation	secular	rio	car	islamic	ethnic	pressure	croats	relatives	evening
english	government	virtue	states	islamic	administration	stations	high	orthodox	air	thursday
wstm	end	iraq	reforms	flights	embassies	official	mph	zagreb	brother	60
ct	taf	kurds	paulo	west	group	million	europe	communism	wife	officials
engl	united	pkk	da	militants	attacks	country	low	saturday	1988	immigrant
ld	week	demirel	cut	minister	indictment	security	taiwan	atrocities	deal	swedish

Fig. 15. *Term Table* depicting the top terms in the topics (30 and 31) from a 30-topic LDA model that are associated with Associated Press (AP) and New York Times (NYT) boilerplate text in the DUC data set.

(AP) and New York Times (NYT) document origins, respectively. The connection between these seemingly dissimilar documents is that they are all AP articles.

Looking at the weights in the *Document Table* images in Figure 14, an interesting pattern emerges. Documents 0–9 (top left image) are articles about Pinochet's arrest. The human-generated cluster labels (red color-coding in the document ID column) show that these documents all belong to this group. LDA has similarly identified this same group of documents as a group, shown by the darker text of the stronger weights in column 44. The weights for documents 0 through 4 in column 30 show a significant connection between these documents and the topic, articles from AP, whereas documents 5 to 9 do not. Unlike document 4, the weighting for document 3 in topic 30 is stronger than its weighting in topic 44 (i.e. document 3 is more strongly aligned with its AP source than with its conceptual content). This same pattern is seen with documents 96, 142, 144, and 193. Our exploration reveals that documents whose conceptual content is outweighed by the source content act as bridges, connecting disparate topics unexpectedly. We identified eleven bridging documents in the DUC data, shown in Figure 16. All are AP articles, and all are

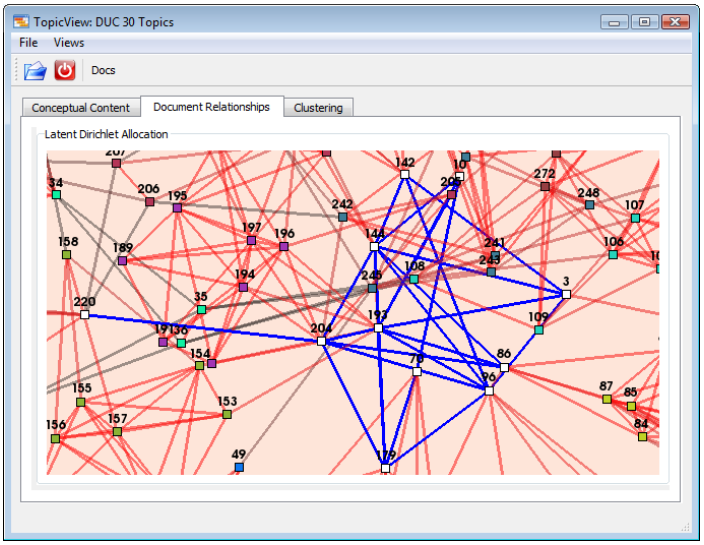


Fig. 16. Bridging documents in the DUC data set highlighted in white with the connecting edges in blue.

short, with some consisting of a single sentence. Comparing AP articles to NYT articles, we find that the NYT articles tend to be longer, sometimes significantly so. Both sets of articles contain tags and header information, but for these short articles, the terms in the headers outnumber the news content.

The bridging documents are in the center of the graph and all of them tend to link with one other, impacting the layout of the connected document groups. The full graph is shown on the lower right of Figure 9. Conveniently, the terms leading to the unwanted edges are easy to isolate and remove. To evaluate the document groupings without the bridging, we reran LDA on just the text from the headlines and story bodies. As expected, the AP and NYT topics disappeared, many of the edges connecting unrelated topics disappeared and document groups became stronger and more visible in the graph, as shown in Figure 17.

Analysis of the NEWSGROUPS data set proceeded in a similar fashion as that of the DUC data set in the case study presented here. Figure 18 presents the *Document Similarity Graph* views for a 20-concept LSA model and 20-topic LDA model. Recall that there are roughly 6 (newsgroup) subject areas represented in this data, and thus a reasonable guess for the number of groupings we could expect to identify in this view could be in the vicinity of 6 as well. And, indeed, even without the nodes being colored by true cluster assignment, we see that there are on the order of 10 or less highly related subgraphs in the LSA document similarity graph. However, in the non-shared layout for the LDA model in the bottom image of Figure 18, there seems to be about 20 smaller, highly related subgraphs that form a spoke-like pattern centered around one larger subgraph. Looking at the shared layout for the LDA model in the top image of the figure, we see that documents across the

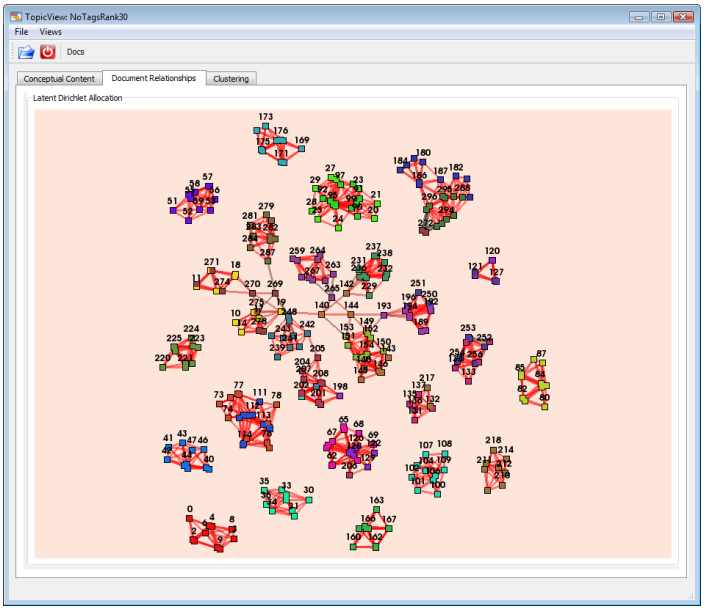


Fig. 17. *Document Similarity Graph* view using a 30-topic LDA model for the DUC data set with documents containing headers and story bodies only.

different clusters appear to be highly interconnected. This behavior resembles that identified for the DUC data and presented in Figure 9.

Inspecting the NEWSGROUPS data set more closely using the views in the *Document Relationships* panel, we can see that the LDA model once again identifies relationships amongst the documents in which there are several documents in the center of the spoke-like structure that are highly interconnected. Figure 19 presents a magnified view of the LDA model similarity graph in which several of these highly interconnected nodes are selected. Looking at the text of the selected individual documents, we again see that most are very short in length. Thus, as with the DUC data, the LDA model is generating a topic to account for the distribution of the terms in these documents across a widely diverse set of subjects. The lack of strong signal identified by the term distributions across these short documents appears to effect the LDA model by using a single topic to attempt to account for the term distributions in those mostly unrelated documents.

Trying to determine how much these short documents impact the document relationships and the structure of the spoke-like LDA model similarity graph, we next turned to the *Clustering* panel views to see how the clustering algorithm groups the highly interconnected documents. Figure 20 presents the clustering assignments when 6 (top) and 10 (bottom) clusters were computed. In the case of 6 clusters, cluster assignments from both the LSA and LDA models are largely mixed, leading to highly heterogeneous clusters (with respect to the true cluster assignments). In the case of 10 clusters, however, the LSA model appears to be able to distinguish

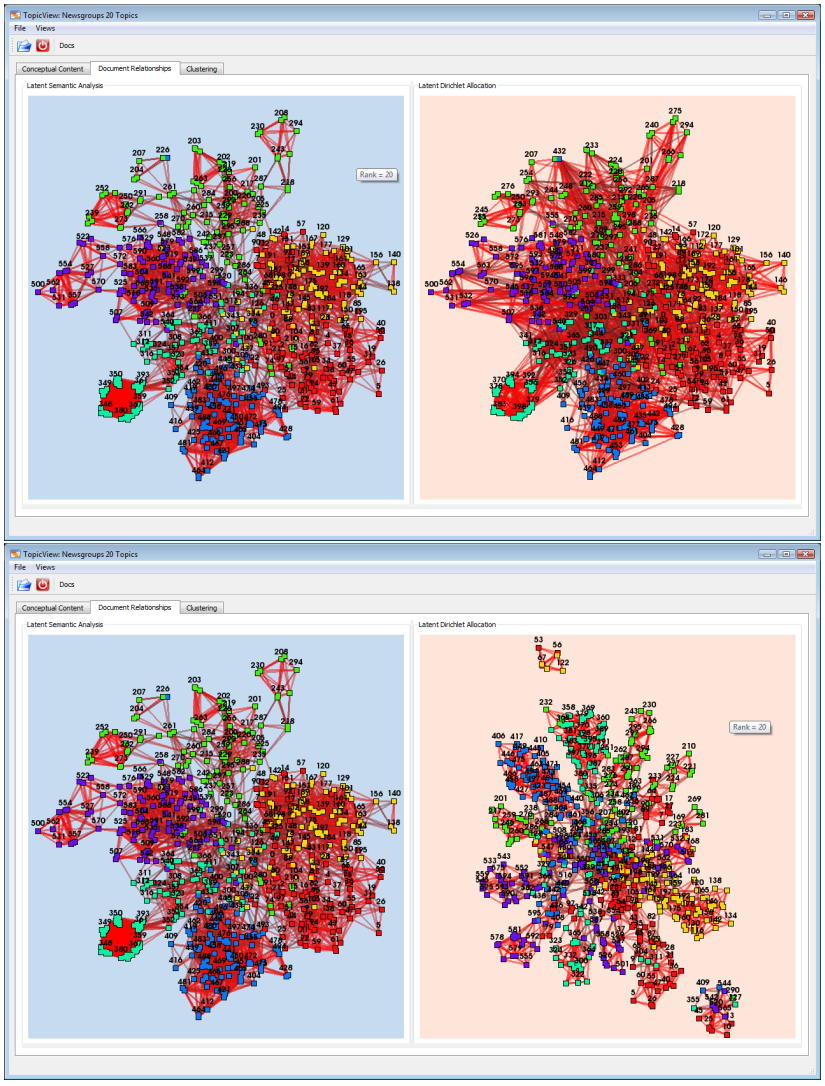


Fig. 18. *Document Similarity Graph* views for the NEWSGROUPS data set with shared node layout for the LSA and LDA models (top) and without shared node layout (bottom).

between the different true clusters, whereas there is still a single large heterogeneous cluster in the *Document Cluster Graph* view for the LDA model.

Further inspection of this large heterogeneous cluster indicates that this cluster contains a majority of the highly interconnected documents. Figure 21 shows the same 10 cluster graphs as in Figure 20 with the large heterogeneous cluster documents selected. We see that the these documents are spread across the various LSA model clusters, evidence that LSA models may not suffer from this bridging or "short-document" behavior, and indication of a major difference between the

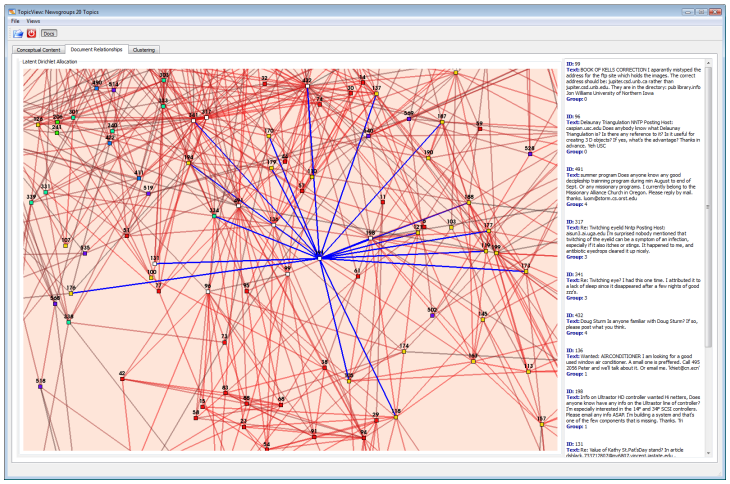


Fig. 19. LDA nodes with large vertex degree selected for the NEWSGROUPS data sets. Document 191 shown with all edges highlighted. Note the short length of the document text for these nodes.

LSA and LDA models. In Figure 22, a magnified view of the cluster is presented along with the text of the documents in that cluster. We see that this provides more evidence that the documents in this cluster are related by their short length more than the topics covered by their sparse text.

Further inspection of this particular cluster in the *Document Relationships* tab indicates that this cluster accounts for most of the relationships in the center of the spoke-like structure of the LDA model document similarity graph. Figure 23 presents the graph views with the heterogeneous cluster documents selected. In the LSA model, those documents are spread across the major subgraphs of documents, whereas for the LDA model, they form a subgraph of their own which effects the interpretation of the relationships amongst the smaller, highly related subgraphs. Figure 24 shows a magnified view of the central high degree vertices, where the vertices are highlighted in white and the edges connecting them are shown in blue. These high degree nodes provide the central set of connections joining the subgraph spokes in the LDA graph. This is further evidence of the differences between the LSA and LDA models, and we see that it greatly impacts both model interpretability and the performance of the clustering algorithm.

7. Conclusions and Further Work

Topic models are the primary means of extracting latent relationships from documents in general text collections, but interpreting them can be a difficult task. Using the TopicView prototype, we were able to refine our ideas for visual model comparison over many iterations, and apply those ideas to a comparison of LSA and LDA models. As our case studies demonstrated, we were able to use visual analytics to explore the models and their impact on analysis tasks, bringing to bear

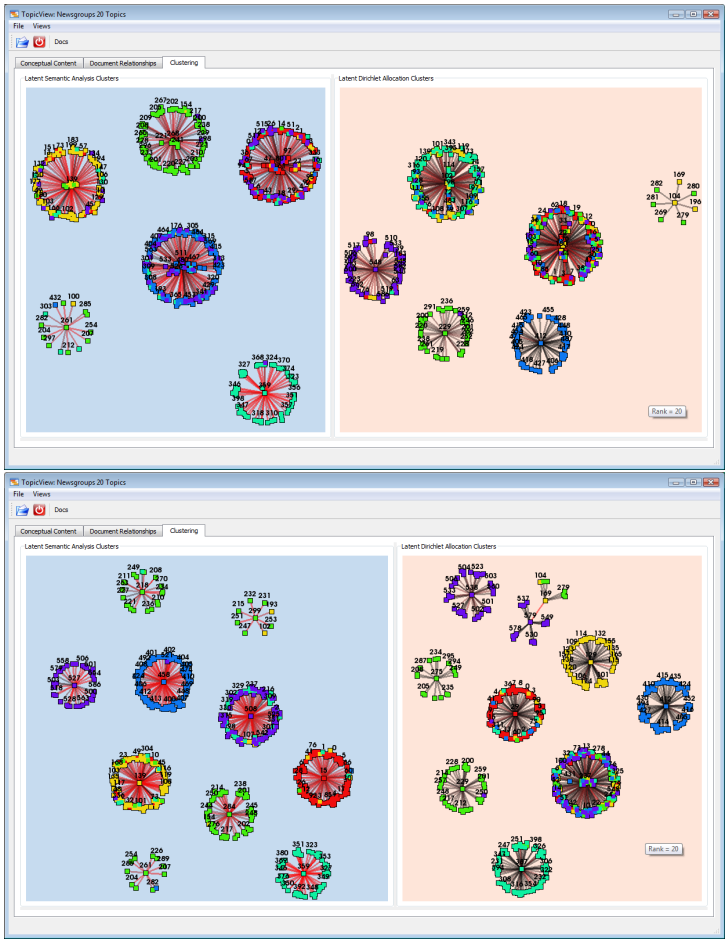


Fig. 20. NEWSGROUPS data set clustering for 20-concept LSA model and a 20- topic LDA models using 6 and 10 clusters.

intuition, subject matter expertise, and specific domain problem solving experience. We were free to explore variations on the models and the impact of those variations efficiently and interactively.

Regarding the specific questions about LSA and LDA that we posed in our introduction, we found that LSA concepts provide good summarizations over broad groups of documents, while LDA topics are focused on smaller groups. LDA's limited document groups and its probabilistic mechanism for determining a topic's top terms support better labeling for document clusters than LSA concepts, but the document relationships defined by the LSA model do not include extraneous connections between disparate topics identified by LDA in our examples.

In future work, we would like to explore other document collections, where clusters share specified amounts of vocabulary overlap, to investigate the generality of

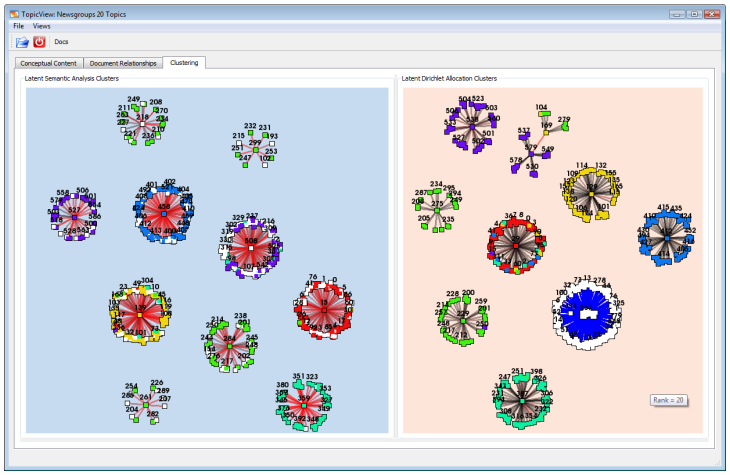


Fig. 21. Selection of documents in a cluster identified using the LDA model and the corresponding documents in the LSA clusters (white nodes). Note that the documents are scattered across all of the LSA model clusters.

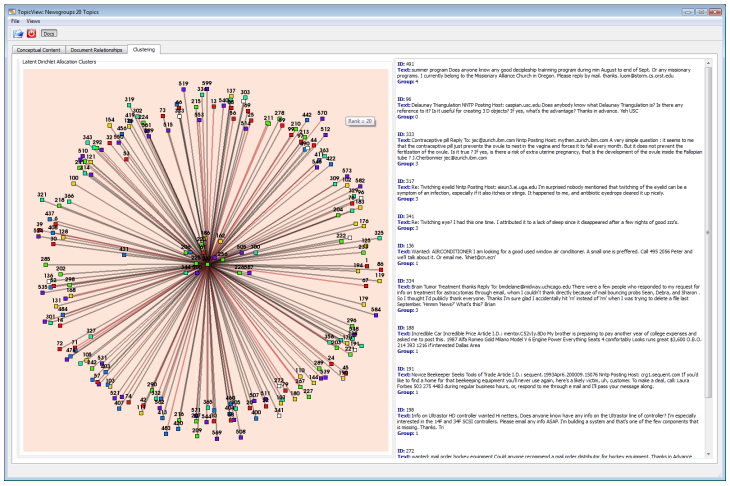


Fig. 22. All but one of the high degree document nodes identified in the NEWSGROUPS data set fall into this single LDA model cluster.

the findings presented in the case studies presented in this paper. With the ALPHABET data set, we can easily do this by varying the size of the vocabulary, the size and number of clusters, and the amount of overlap between documents within and across clusters. We would further like to explore additional document modeling methods, including nonnegative matrix factorizations (NMF)[27] and extensions to LDA that have shown improved performance in document clustering applications, such as mixture of von Mises-Fisher models.[28]

More broadly, we envision the use of tools like TopicView as an initial step in

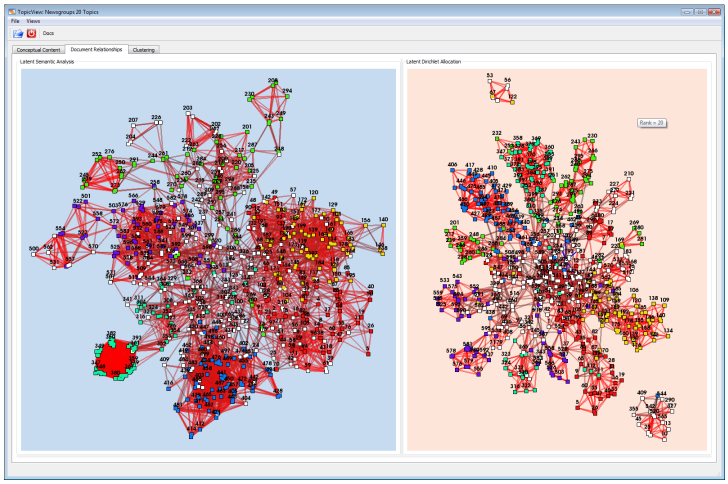


Fig. 23. *Document Similarity Graph* views of the NEWSGROUPS data set, with high degree nodes selected.

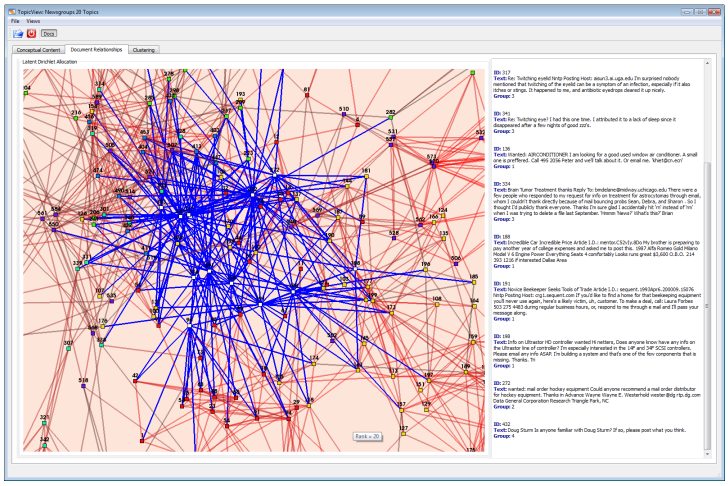


Fig. 24. A magnified LDA *Document Similarity Graph* view of the bridging nodes in the NEWSGROUPS data set, highlighting the "short-document" nodes and the edges connecting them.

exploring very large data sets. Users often apply analysis methods in arbitrary ways, leading to suspect conclusions that may not be truly supported by their data. We imagine users applying the visual analytics techniques developed for TopicView to understand small, annotated samples of a larger collection, allowing them to make better informed decisions about which topic model(s) can best facilitate analysis for the larger whole. Of course, further work will be necessary to demonstrate that visual analytics can be generalized in this fashion.

Acknowledgments

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

References

1. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.
2. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, March 2003.
3. V. Crow, K. Pennock, M. Pottier, A. Schur, J. Thomas, J. Wise, D. Lantrip, T. Fiegel, C. Struble, and J. York. Multidimensional visualization and browsing for intelligence analysis. In *Proc. GVIZ*, September 1994.
4. C. Chen. Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST*, 57(3):359–377, December 2005.
5. J. A. Wise. The ecological approach to text visualization. *JASIST*, 50(13):1224 – 1233, 1999.
6. G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, and B. N. Wylie. Knowledge mining with VxInsight: Discovery through interaction. *Journal of Intelligent Information Systems*, 11(3):259–285, 1998.
7. T. Kakkonen, N. Myller, E. Sutinen, and J. Timonen. Comparison of Dimension Reduction Methods for Automated Essay Grading. *Educational Technology & Society*, 11(3):275–288, 2008.
8. I. Biro, A. Benczur, J. Szabo, and A. Maguitman. A comparative analysis of latent variable models for web page classification. In *Proceedings of the 2008 Latin American Web Conference*, LA-WEB '08, pages 23–28, Washington, DC, USA, 2008. IEEE Computer Society.
9. T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in semantic representation. *Psychological Review*, 114(2):211–244, April 2007.
10. C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Proc. VAST*, pages 91 –98, October 2009.
11. M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.
12. T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(Suppl. 1):5228–5235, April 2004.
13. J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.
14. M. M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, 2009.
15. L. Kaufman and P. Rousseeuw. *Clustering by Means of Medoids*. Reports of the Faculty of Mathematics and Informatics. Delft University of Technology. Fac., Univ., 1987.
16. I. Katsavounidis, C. C. Jay Kuo, and Z. Zhang. A new initialization technique for generalized Lloyd iteration. *IEEE Signal Processing Letters*, 1(10):144–146, October 1994.
17. Review, Comparative, J. He, M. Lan, C. lim Tan, S. yuan Sung, and H. boon Low.

Initialization of cluster refinement algorithms:. In *in Proceedings of International Joint Conference on Neural Networks (IJCNN*, pages 297–302, 2004.
18. A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 6–17, 2002.
19. C. Hennig. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1):258–271, September 2007.
20. D. J. Hand. Data clustering: Theory, algorithms, and applications by guojun gan, chaoqun ma, jianhong wu. *International Statistical Review*, 76(1):141–141, 2008.
21. 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.
22. K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.
23. D. M. Dunlavy, T. M. Shead, and E. T. Stanton. ParaText: Scalable text modeling and analysis. In *Proc. HPDC*, pages 344–347, 2010.
24. B. Wylie and J. Baumes. A unified toolkit for information and scientific visualization. In *Proc. Visualization and Data Analysis*, volume 7243, page 72430H. SPIE, 2009.
25. P. J. Crossno, A. T. Wilson, D. M. Dunlavy, and T. M. Shead. Topicview: Understanding document relationships using latent dirichlet allocation models. In *Proc. IEEE Workshop on Interactive Visual Text Analytics for Decision Making*. IEEE, 2011.
26. C. Ziemkiewicz and R. Kosara. Laws of attraction: From perceptual forces to conceptual similarity. *IEEE Transactions on Visualization and Computer Graphics*, 16:1009–1016, November 2010.
27. M. Berry and M. Browne. Email surveillance using non-negative matrix factorization. *Comput. Math. Organ. Th.*, 11(3):249–264, October 2005.
28. J. Reisinger, A. Waters, B. Silverthorn, and R. J. Mooney. Spherical topic models. In *Proc. ICML*, pages 903–910, 2010.