SAND2011-2182C

# TopicView: Visually Comparing Semantic Models

Patricia J. Crossno, *Member, IEEE*, Andrew T. Wilson, Daniel M. Dunlavy, and Timothy M. Shead

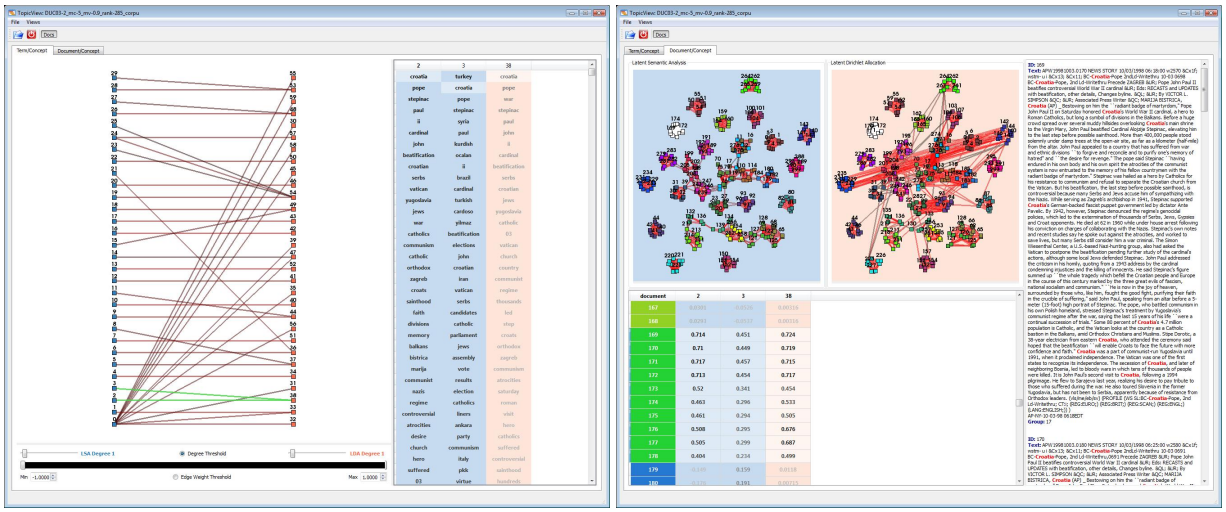

Fig. 1: The TopicView user interface. At left, conceptual content with concepts and terms from LSA in blue and topics and terms from LDA in red. The image on the right displays document relationship views with the similarity graph for LSA on the left and the similarity graph for LDA on the right.

**Abstract**—We present TopicView, an application for visually comparing and exploring models of text corpora created using Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). TopicView uses multiple linked views to visually analyze the conceptual content and the document relationships in the models generated by each algorithm. Conceptual content is compared through the combination of (i) a bipartite graph matching LSA concepts with LDA topics based on their cosine similarities 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 any documents selected in either of the graphs or the table. We demonstrate the utility of TopicView's visual approach by comparing the LSA and LDA models for two example corpora. Using the tool, we have found that LSA concepts provide summarizations over broad groups of documents, while LDA topics are focused on smaller groups of documents. 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. On the other hand, the document relationships defined by the LSA model do not include the extraneous connections between disparate topics shown by the LDA document similarity graph for one of our examples. TopicView reveals that these extra edges occur only when the document is short enough that the common terms relating to the source (e.g., newswire headers) create a stronger signal than the main document content (e.g., newswire story).

**Index Terms**—text analysis, latent semantic analysis, latent dirichlet allocation.

## 1 Introduction

Latent Semantic Analysis (LSA) [11] and Latent Dirichlet Allocation (LDA) [4] are two popular mathematical approaches to text analysis. Both LSA and LDA have been implemented as part of the Titan Informatics Toolkit [8][21], an open source project developed by Sandia National Laboratories and Kitware, Inc. The work described in this paper was motivated by questions posed by application developers: How closely do LSA's concepts correspond to LDA's topics? How similar are the most significant terms in LSA concepts to the top terms of corresponding LDA topics? Are the same documents affiliated with matching concepts and topics? Do the document similarity graphs produced by the two algorithms contain similar document clusters? How well do document clusters found in their respective similarity graphs match human-generated clusters?

LSA and LDA have much in common. They both (i) use bag-of-words modelling, (ii) start by transforming text corpora into term-document frequency matrices, (iii) require as input the number of concepts or topics that the algorithm will generate, (iv) produce weighted term lists for each concept or topic, (v) produce concept or topic content weights for each document, and (vi) produce outputs that can be used to generate document similarity graphs. Yet despite these similarities, the two algorithms generate very different models. LSA uses singular value decomposition (SVD) to project documents into a shared semantic space, in which similar documents are positioned near one another. LDA uses a Bayesian model that treats each document as a mixture of latent underlying topics, where each topic is modelled as a mixture of word probabilities from a vocabulary. Furthermore, although LSA and LDA outputs can be used in similar ways by applications, their output values represent entirely different quantities, with different ranges and meanings. LSA produces term-concept and document-concept correlation matrices, where correlation values range between -1 and 1 with negative values indicating inverse correlations. LDA produces term-topic and document-topic probability

- *Patricia J. Crossno, Andrew T. Wilson, Daniel M. Dunlavy, and Timothy M. Shead are with Sandia National Laboratories, E-mail: {pjcross, atwilso, dmdunla, tshead}@sandia.gov.*

*Manuscript received 31 March 2010; accepted 1 August 2010; posted online 24 October 2010; mailed on 16 October 2010.*
*For information on obtaining reprints of this article, please send email to: tvcg@computer.org.*

matrices, where probabilities range between 0 and 1. How does one numerically compare correlations with probabilities?

Our approach is to move away from statistical comparisons and instead to focus on human consumable differences. Although visual analytics applications can use a variety of metaphors for visualizing a document collection, such as scatter plots [9], graphs [6], or landscapes [20], they all rely on document similarity measures to position similar documents near one another in the visualization. These representations are often combined with labels to identify the topical or conceptual contents of document groups [10]. Consequently, we focus our comparison on the document relationships and conceptual categories identified by LSA and LDA for a given corpus.

In this paper we present *TopicView*, an application designed to visually compare and interactively explore LSA and LDA models from this user-based perspective. As shown in Figure 1, tabbed sets of linked views compare conceptual content (left image) and document relationships (right image), both between individual documents and between documents and conceptual content. In addition to describing the design and implementation of TopicView, we also present our insights from using TopicView to contrast the models generated by each algorithm for two small corpora.

## 2 Related Work

Griffiths et al. compare semantic spaces models with generative probabilistic topic models from the perspective of each approach's ability to model human semantic memory [15]. In particular, they are concerned with the ability of a model 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, our work focuses on comparing the impact that model differences have on visual analytics applications, using visualization to do the comparison.

Collins et al. combine tag clouds with parallel coordinates to form Parallel Tag Clouds[7], an approach for comparatively visualizing differentiating words within different dimensions of a text corpus. Word lists are alphabetical, with word size used to show word weighting. Similar to parallel coordinates, matching terms are connected across columns. Although we have similar goals in comparing term lists, we feel that our approach of sorting terms combined with scaling text luminance by weight provides a clear comparison of the relative significance of terms across concepts and topics without the layout complications and potential overlaps encountered when words are drawn at vastly different scales.

## 3 Algorithm Overviews

To introduce some of the algorithm specific terminology, we provide a brief overview of the text analysis algorithms.

### 3.1 Latent Semantic Analysis

LSA computes a truncated SVD of a term-document matrix [3], i.e., the collection of feature vectors associated with the documents in a text collection, or corpus. More specifically, the rank-$k$ LSA model of a term-document matrix, $A \in \mathbb{R}^{m \times n}$, 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. 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 $k \ll \min(m,n)$, leading to a reduction in model noise and computation for many analysis methods.

We compute distances, or similarity scores, between all pairs of documents using cosine similarities, defined as

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

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

**for** $k = 1$ to $K$ **do**
  Draw $\phi^k \sim \text{Dirichlet}(\beta)$
**end for**
**for** $d = 1$ to $D$ **do**
  Draw $\theta \sim \text{Dirichlet}(\alpha)$
  Draw $N \sim \text{Poisson}(\xi)$
  **for** $i = 1$ to $N$ **do**
    Draw $z \sim \text{Multinomial}(\theta)$
    Draw $w \sim \text{Multinomial}(\phi^{(z)})$
  **end for**
**end for**

Algorithm 1: Generative algorithm for LDA. This will generate $D$ documents with $N$ tokens each. Each token is drawn from one of $K$ topics. The distributions over topics and terms have Dirichlet hyperparameters $\alpha$ and $\beta$ respectively. The Poisson distribution over the token count may be replaced with any other convenient distribution.

### 3.2 Latent Dirichlet Allocation

The LDA algorithm [14] builds upon a generative statistical model for documents shown in Algorithm 1. Given a few parameters – a vocabulary of $W$ distinct words, a number of topics $K$, two smoothing parameters $\alpha$ and $\beta$ and a prior distribution over document lengths (typically Poisson) – this generative model creates random documents whose contents are a mixture of topics.

In order to use LDA to model the topics in an existing corpus we must invert the generative model and infer model parameters from the data instead of generating documents from the parameterized model. Specifically, for a corpus containing $D$ documents we want to learn $\phi$, the $KxW$ matrix of topics, and $\theta$, the $DxK$ matrix of topic weights for each document. The remaining parameters $\alpha, \beta$ and $K$ are specified by the user. The Titan implementation of LDA uses collapsed Gibbs sampling [13, 5] to estimate the $\theta$ and $\phi$ matrices.

## 4 TopicView

TopicView loads a small corpus and uses both LSA and LDA to generate models of the data. A shared pre-processing stage produces a term-document matrix and a term dictionary that serve as inputs to both algorithms. Identical rank (LSA) and topic counts (LDA) are used to generate matching numbers of concepts or topics in their respective outputs. The rank/topic count is input by the user. Outputs are run through the same cosine similarity, edge threshold, and graph layout filters to produce document similarity graphs. Our goal throughout this process is to limit the differences to be just those differences attributable to the two algorithms. While this glosses over some differences between the algorithms (especially singular vectors vs. probability distributions), we feel that this represents the path of least uncertainty for the goals of our study.

Conceptual content and document relationships are visualized on separate tabbed displays. The views are designed to enable exploration of progressively more detailed relationships from the corpus level down to reading individual documents. Color coding is used to differentiate the models. LSA-generated components are in blue and are positioned on the left (left nodes in the bipartite graph, left-most columns in the tables, and the left document similarity graph), while LDA-generated model components are in red and are on the right. We

have picked a set of graduated colors from ColorBrewer's [16] single hue sequences to provide a family of colors that can be unambiguously identified with each algorithm.

Throughout this paper we use the combined term *concepts/topics* to refer to the generic idea of the conceptual content generated by the algorithms. We do this because (i) the literature for each algorithm has a slightly different vocabulary for describing the conceptual content produced and we want to preserve the earlier terminology in our descriptions (ii) the numeric values represent different types of quantities with LSA producing correlations and LDA generating probabilities (iii) our analysis of LSA and LDA has shown that although concepts/topics have surface similarities, they are not interchangeable and do have important differences.

### 4.1 Conceptual Content

To answer our motivating questions regarding the differences between LSA and LDA, we need to compare concepts and topics. We do this through a combination of linked views. At the highest level, we want to know how concepts compare to topics, without getting into the details of the ideas represented by either one. A bipartite graph provides an abstract overview of these relationships by connecting concepts and topics with weighted edges. On a lower level, conceptual content is represented by the relative strengths of the terms within each concept/topic. Although we cannot directly compare the weightings assigned to individual terms between concepts and topics, we can visually compare the order and relative weighting of their terms.

#### 4.1.1 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. 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 by the strength of the correlation (see Figure 1, left side of left image).

To calculate the concept and topic similarities, we first scale LSA's left singular vector output matrix by its singular values, then concatenate the result with the transpose of LDA's $\phi$ matrix. After computing the cosine similarities (using Eq. 2) of the concatenated matrix, we truncate the similarity matrix to be just the upper right quadrant, retaining only unique similarities between LSA concepts and LDA topics (the discarded quadrants contain either similarities within concepts, similarities within topics, or the reverse edges for the similarities in the upper quadrant). The truncated edge list is sorted in descending order. The edge weights are correlations ranging in value from -1.0 to 1.0, which we color-code from blue to black to red to preserve the distinction between negative and positive correlations. Using bright red for the strongest positive correlations allows us to easily find the strongest concept/topic pairings. Although inverse correlations can also be discovered, we cannot say much more about these links than that those concept/topic pairs have nothing in common.

With LSA's truncated SVD model, each concept captures the maximum variation in the data, which is then subtracted from the results prior to computing the next concept. Consequently, LSA's concept order conveys information about the variation in each rank, which we preserve in the bipartite graph layout by fixing the LSA node positions in their rank order. We use a greedy approach for placing LDA nodes relative to the LSA nodes in order to draw the strongest correlations with horizontal edges. Using the sorted edge list, we step through the list in order of declining edge strength. If the LDA node associated with an edge has not yet been placed, we position it horizontally across from the LSA node for that edge. As each LDA node is placed, it is flagged as "used" and any further edges that use that LDA node are ignored and do not factor into the layout.

We provide two interactive filtering mechanisms for reducing the number of bipartitie graph edges: *Degree Threshold* and *Edge Weight Threshold*, the controls for which are visible below the graph in Figure 1. *Degree Threshold* independently controls the minimum vertex degree for each side of the bipartite graph with a separate slider. To fulfill the degree of a particular vertex, edges are drawn in descending weight order so that the strongest edges are seen first. Although the sliders control the minimum number of edges coming from each node, some nodes (such as LSA node 0) may exceed that degree because the node is on the receiving end of another node's minimum edge count. Alternatively, *Edge Weight Threshold* acts to display all of the edges whose weights fall within a user specified range. The range can be input using either a double-ended slider or independent specification of *Min* and *Max* values.

Nodes and edges in the bipartite graph are selectable. Selecting an edge is equivalent to selecting its two end nodes. Nodes correspond to columns in the Term Table and in the Document Table on the Document Relationships tab. Selection reduces the columns displayed in both tables to be just the selected concepts and topics. This provides a column adjacency that is useful for comparing word lists in the Term Table or looking to see which documents contribute most heavily to those concepts/topics in the Document Table. Clicking anywhere outside the graph clears the current selection and restores the entire set of columns.

#### 4.1.2 Term Table

The *Term Table* presents the terms for each concept or topic sorted in decreasing order of importance. Text color is used to provide an additional cue about the relative weights of terms, varying from black for the most highly weighted terms to a light gray (192/255) for the lowest weighted terms. We do not use the full luminance range because we found that white text was illegible against the pale blue and pink backgrounds. Background colors are in turn constrained to be light enough to provide sufficient contrast with the text and graph components in the various views, but dark enough to permit at least one additional level of lightness for highlighting.

Since we are most interested in distinguishing weighting differences at the high end of the scale, we spread this part of the range by using a logrithmic mapping that increases the number of luminance steps as we approach black. Although text in both LSA and LDA columns use a copy of this same logarithmic lookup table, each lookup table's range is independently scaled to match that algorithm's range of values. Consequently, luminance differences can only be directly compared within the same algorithm.

Individual terms 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, where every instance of that term within the selected documents is written in red.

### 4.2 Document Relationships

Document clustering as shown through *Document Similarity Graphs* provides an alternative view of LSA and LDA model differences. We informally define a cluster as a group of documents with strong links between members of the group and weak links outside the group. Although there is a tendency to try to identify concepts or topics with clusters, the weightings shown in the *Document Table* demonstrate that document groups frequently contribute in varying degrees to multiple concepts or topics (weightings spread across rows). Similarly, concepts and topics typically include multiple document groups (weightings spread across columns). The visual combination of the graphs and tables on the same tab enables the user to locate and select the documents associated with either conceptual content or clusters and then read their full texts in the *Document Text* view.

#### 4.2.1 Document Similarity Graphs

To compute document similarity graphs for LSA we calculate cosine similarities using the right singular vectors, scaled by the singular values. For LDA we compute cosine similarities using the $\theta$ matrix. Cosine similarities (computed using Eq. 2) compute edge weights between every pair of documents. Consequently, we need to reduce the visual clutter by thresholding the edges. 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 .9) (iii) if the number of highly weighted edges for that 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 layout the graphs using a linear time force-directed layout algorithm. Although both graphs can be independently laid out based on their individual edges, we provide an option to view both graphs using a shared layout as shown in Fig. 6. In the shared layout, the graphs have identical node placement (taken from the LSA layout because it tends to cluster better), and the differences are limited to the edges.

Each document is labeled with its document ID and color-coded by a ground-truth category. Edges are color-coded using saturation to indicate similarity weights, with low values in gray and high values in red. Nodes and edges can be selected either individually, or rectangular rubber-banding can be used to select everything within a region. Selected nodes and edges are drawn in white, as that is the only color that is sufficiently unique to stand out from the group labels. The two graphs are linked, so corresponding graph nodes and edges are selected in both (note that some edges may exist in one graph and not the other). The selected documents are also highlighted in the *Document Table* and their full text is displayed in the *Document Text* view.

#### 4.2.2 Document Table

The *Document Table* is the concatenation of the transpose of LSA's right singular vector, scaled by its singular values, with LDA's $\theta$ 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 for darker, more highly weighted documents within a concept or topic.

The visible columns within the table are controlled by selecting nodes in the *Bipartite Graph*. Subsetting the column display facilitates side-by-side comparisons of the relative weightings of the most significant documents associated with a set of concepts or topics. It can be used to formulate hypotheses about the relationship between conceptual content and specific documents. Selecting rows within the table will highlight nodes in both graphs and display the selected document centents in the *Document Text* view.

#### 4.2.3 Document Text

This view displays the full text contents of multiple documents, selected either through the similarity graphs or the *Document Table*. Each document is displayed as three fields: the document ID, the raw text of the document, and the categorical ID. If a term is selected in the *Term Table*, that term is highlighted in red throughout the raw text. For longer text selections, the view can be scrolled.

### 4.3 Case Studies

In this section, we present the results of using TopicView to find similarities and differences between LSA concepts and LDA topics generated from synthetic and real-world document collections. The goals of the case studies presented here include the following:

- Illustrate the use of TopicView for efficient and effective navigation of relationships between LSA concepts and LDA topics.
- Determine the relationship between LSA concepts and LDA topics with respect to the most important terms and overall term distributions associated with topically related clusters of documents.
- Identify strengths of the different modeling techniques (i.e., LSA and LDA) with respect to document clustering and model interpretability.

A detailed description of the LSA implementation in TopicView can be found in [12]. In the following case studies, $\alpha$ is set to 1 when scaling by the singular values. The LDA implementation inputs two parameters (in addition to the number of topics), the number of sampling iterations and the number of burn in iterations, which have been set to 1 and 200, respectively.

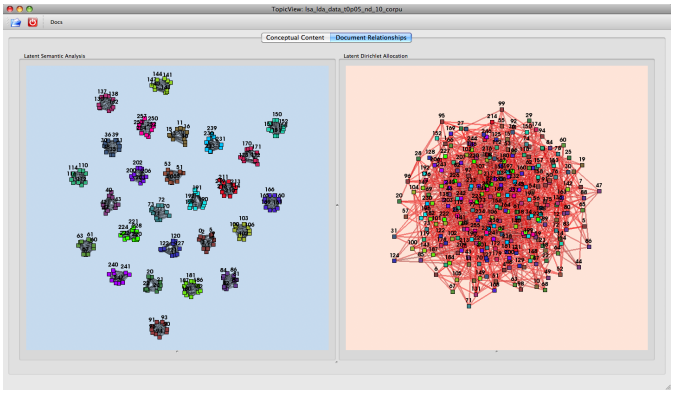


Fig. 2: Graphs depicting document relationships modeled using LSA (left) and LDA (right) for the *alphabet* data set.

#### 4.3.1 Data

Two data sets were used in the case studies. The initial case study uses an artificial document set, the *alphabetic* data set, in which the terms in each cluster are entirely disjoint from one another. The second case study uses the *DUC* data set, a real-world document collection in which terms and concept/topics do overlap. Both data sets have human-generated cluster labels, which are used to color-code the document groups to help answer the question of how well the algorithms' clusters match human-generated clusters.

The *alphabetic* data set consists of 26 clusters containing 10 documents each, where each cluster consists of documents made up exclusively of terms starting with the same letter. For example, the first cluster contains documents consisting of terms starting with the letter "a", the next cluster consists of documents containing only terms starting with the letter "b", and so on. The term set was constructed starting from a dictionary of 1000 words staring 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 consistiing of 26000 terms. For each of the 10 documents from each cluster, 100 terms were sampled with replacement from 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 can be easily identified visually, as all terms associated with a cluster start with the same letter.

The *DUC* data set is a collection of newswire documents from the Associated Press and New York Times that were used in the 2003 Document Understanding Conference (DUC) for evaluating document summarization systems [18]. The collection consists of 298 documents categorized into 30 clusters, with each cluster containing roughly 10 documents focused on a particular topic or event.

#### 4.3.2 Case Study Using Alphabet Data

In this study, we work with the *alphabet* data set to illustrate relationships of LSA and LDA modeling applied to a collection that contains clusters of documents that are independent with respect to term usage. As LSA concepts are, by definition, orthogonal latent feature vectors, we expect that LSA should be able to model each of the document clusters using a single concept (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. However, the purpose of this study is to illustrate the use of TopicView to identify differences in LSA and LDA models when one model is able to exactly match the data.

Figure 2 presents TopicView's *Document Similarity Graphs* for the LSA (left) and LDA (right) models applied to the *alphabet* data set. From this view, we see that the LSA model clusters the *alphabet* data set well; there are 26 disconnected components in the graph, and each

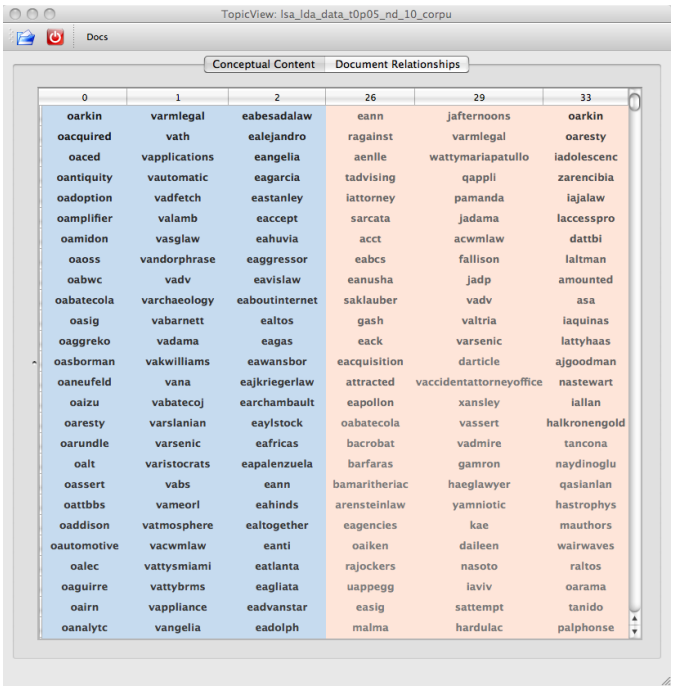


Fig. 3: Relationships between LSA concepts (blue nodes) and LDA topics (red nodes) for the *alphabet* data set. The left side of the screen illustrates that LSA concepts identify the independent terms sets in the data, whereas terms starting with different letters are highly mixed across LDA topics.

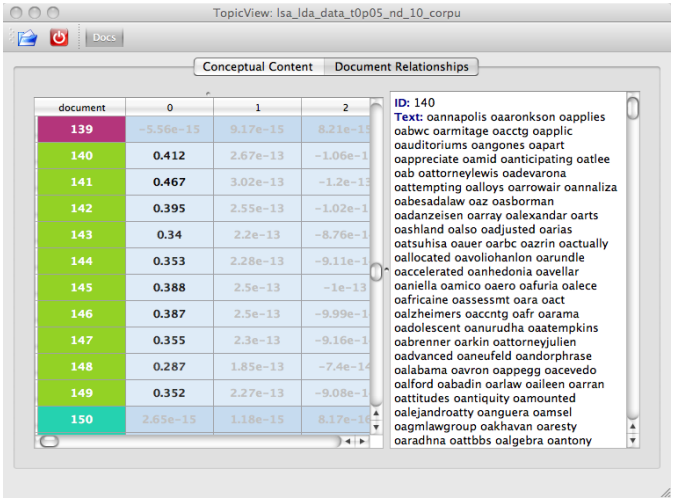


Fig. 4: Document Table and Document Text views in the Document Relationship Panel illustrate that LSA concept 0 contains only "o" documents from the *alphabet* data set.

component consists of nodes colored with the same cluster label. On the other hand, the LDA model is unable to partition the data, indicating strong relationships between all documents across the entire collection. The *Term Table* on the Conceptual Content tab, as shown in Figure 3, can be used to better understand the model differences with respect to the 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 (i.e., topics 33, 29, and 26, respectively). We can see that words beginning with "o", "v", and "e" are those most highly correlated with LSA 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, respectively, but in general there is no clear connection to any particular document cluster. Further investigation using the Document Table and Document Text views in the Document Relationship tab confirm that LSA models the clusters correctly; see Figure 4 for 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 is modeling the clusters accurately and LDA is not, we used TopicView's Bipartite Graph view to identify relationships between the LSA concepts and LDA topics. Figure 5 shows the Bipartite Graph view 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., columns of $\phi$), 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 the LSA concepts and LDA topics are weak as indicated by the gray colors of edges (as opposed to bright red edges that indicate very strong relationships). The image on the left 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 not at all related to the LSA concepts, each of which models one of the document clusters. Furthermore, by exploring the relationships using the edge

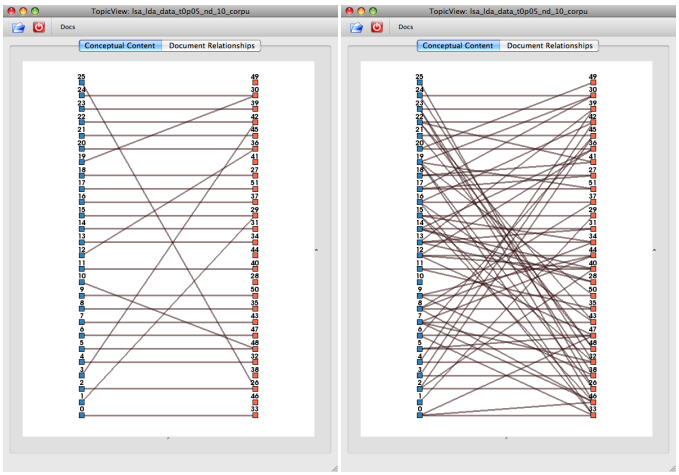


Fig. 5: Graphs depicting conceptual contents relationships for the *alphabet* data set modeled with LSA (blue nodes) and LDA (red nodes) using a minimum degree threshold of 1 for LSA concept nodes (left) and a minimum edge threshold of 0.1451 (right)

weight thresholding controls (right image in the figure), we see that at a threshold of 0.1451 (found easily using the threshold slider controls) 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 capturing the terms relationships within each document cluster.

These outcomes are consistent with our expectations for this synthetic data set. 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, these methods rely on co-occurrence of terms between documents to guide a random walk toward more probable topic configurations. In the *alphabet* data set we explicitly suppress term co-occurrence between clusters. 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, we expect the topics from LDA to be random with more or less uniform character. This is indeed what we observe.

This case study illustrates how TopicView can be used to explore relationships between LSA concepts and LDA topics with respect to term distributions and document clustering. The examples here indicate that LDA may not be best suited for clustering document collections with (little or) no mixing of terms across clusters (i.e., independent document clusters). In the next case study involving a real-world collection of documents, we see the typical performance of LSA and LDA models as reported in the literature and show how TopicView can be used to explore the similarities and differences between these two techniques.

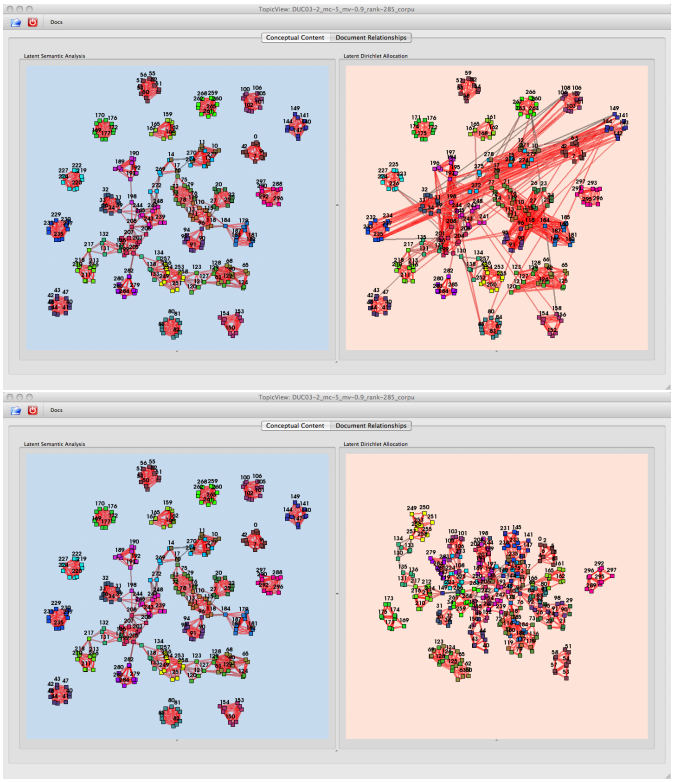


Fig. 6: Graphs depicting document relationships modeled using LSA (left) and LDA (right) for the *DUC* data set. The top set of images is using the shared layout and the bottom set is using the layouts independently computed from the different models.

### 4.3.3 Case Study Using DUC Data

In the previous section, we illustrated how TopicView can be used to investigate and explore relationships between LSA and LDA models. However, that case study was designed solely to illustrate the use of the different components in TopicView applied to a problem in which dramatic differences between LSA and LDA models would exist. However, collections of documents with topic clusters containing no overlap in vocabulary, as in the case of the *alphabet* data set, do not appear often in real-world analysis applications. Even when document collections contain clusters of documents across a wide range of disparate subject areas, there is a large degree of overlap in vocabulary across documents in different clusters. To investigate the relationships between LSA and LDA modeling on such a real-world collection of documents, we applied TopicView to the *DUC* data set. Although the *DUC* data set contains subsets of document whose general topics are very different to the annotators, we show how TopicView can be used to explore how LSA and LDA models are similar in identifying clusters with consistent term distributions across documents in particular clusters, but also different in how weak connections between document clusters are modeled.

As in the previous case study, we start our analysis by exploring the relationships between the LSA and LDA models examining the document similarity graphs for the two models. Figure 6 shows the view of the document similarity graphs sharing the LSA layout (top) and the view where each model's layout is used (bottom). We see that the LSA model (left graphs) results in a graph with 14 disconnected components, 13 of which are clusters identified by the human annotators. The larger component located in the center of the layout indicates that many of the clusters are related in their term distributions to some degree. We see that there are many subgraphs which correspond to true document clusters as shown by the node colorings in the largest component that are connected 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 (top image in Figure 6), we see that there are many inter-cluster relationships identified. This indicates that LSA and LDA are clearly modeling different characteristics in the data. When the LDA-specific layout is viewed, we get a much better sense of the clustering produced by the LDA model; there are disconnected components (indicating tight document clusters) and a similar subgraph structure as the LSA model. Thus, we conclude that the LDA model also provides a useful clustering. However, TopicView can be used to visually explore in more detail the similarities and differences between these difference clusterings.

Figure 7 illustrates how TopicView's Conceptual Content Bipartite Graph views can be used to indicate relationships between the LSA concepts and LDA topics. Using either 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 is modeling the statistical variance of terms across the documents and thus acts as a generic concept that summarizes all of the main interactions between documents as a functions of the terms appearing in those documents [11].

Exploring beyond this unique relationship of LSA concept 0, we also see several cases where multiple LSA concepts are strongly connected to a single LDA topic or vice versa. Two such examples include (a) LSA concepts 9 and 11 being strongly connected to LDA topic 44, and (b) LSA concepts 6 and 21 being strongly connected to LDA topic 36. Note that case (a) is presented in Figure 1, where all of TopicView's panels and views are shown. The top 10 terms associated with the LSA concepts and LDA topics that are part of the relationships in cases (a) and (b) are shown in the top and bottom images in Figure 8, respectively. 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

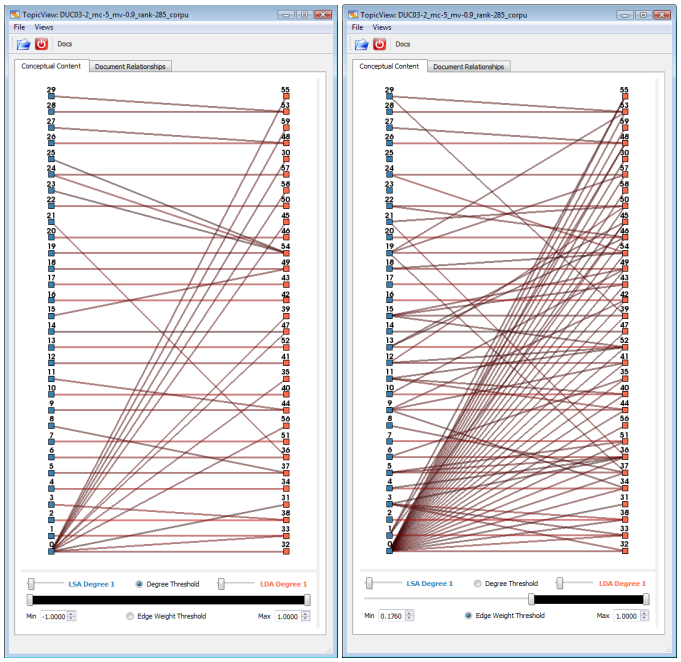

Fig. 7: Graphs depicting conceptual content relationships for the *DUC* data set modeled with LSA (blue nodes) and LDA (red nodes) using a minimum degree threshold of 1 for LSA/LDA nodes (left) and a minimum edge threshold of 0.1760 (right).

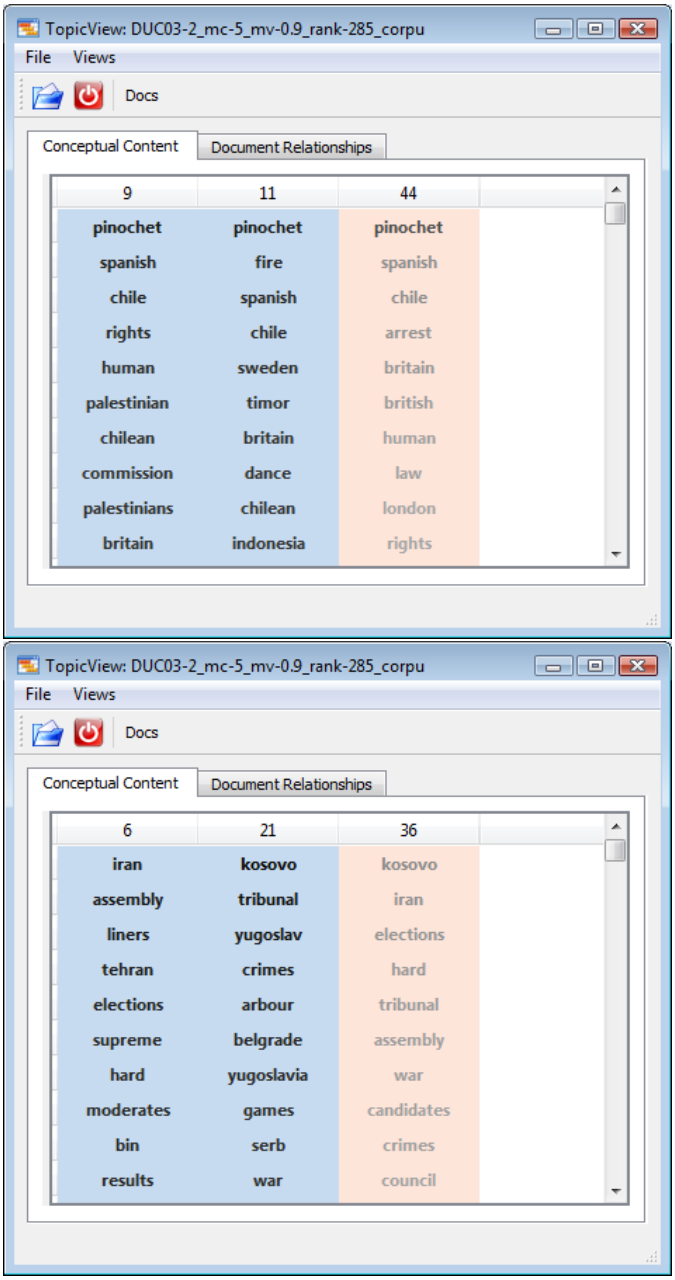


Fig. 8: Top 10 terms associated with the concepts and topics where multiple LDA concepts are strongly connected to a single LDA topic.

Pinochet cluster with clusters of documents about a dance hall fire in Sweden and poitical unrest in Timor, which do not appear related. Using TopicView's 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 document 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 (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 that account 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 has been discussed in the [4], 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 both the Pinochet document cluster well with concept 9 and the more subtle, polysemous cross-cluster term relationships between the clusters regarding 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, we see 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 combined use of the Document Table and Document Text views to identify term relationships leading to the combined LDA topic.

We conclude from this case that LSA and LDA model the most tightly coupled document clusters well (as indicated by many strong horizontal edges in Figure 7), but model more subtle relationships between documents and thus clusters in different ways. By using TopicView, we were able to quickly identify and explore these differences. Moreover, using TopicView, we were able to specifically identify the documents and terms that led to the differences in the LSA and LDA

Sandia National Laboratories
National Nuclear Security Administration

models.

## 5 Conclusions and Future Work

We showed that TopicView helps identify similarities and differences between different text modeling methods. We have provided initial evidence in support of a multiple-model text modeling approach, that when coupled with capabilities available in TopicView, will help analysts investigate a documents and collections of documents from multiple (and perhaps related) perspectives.

In our future work, we would like to explore other document collections, where clusters share more or less vocabulary overlap to investigate how general are the findings presented in the two case studies in this paper. With the alphabet framework, we can easily do this. We can explore relationships between different document modeling methods, such as nonnegative matrix factorizations (NMF) [2] and extensions to LDA that have shown improved performance in document clustering applications, such as mixture of von Mises-Fisher (moVMF) models [1, 19].

## Acknowledgments

This work was funded by the Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

## References

[1] A. Banerjee and S. Basu. Topic models over text streams: A study of batch and online unsupervised learning. In *Proc. SIAM Intl. Conf. on Data Mining*. SIAM, 2007.

[2] M. Berry and M. Browne. Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, 11(3):249–264, Oct. 2005.

[3] 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.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[5] S. Chatterji and L. Pachter. Multiple organism gene finding by collapsed gibbs sampling. In *RECOMB '04: Proceedings of the eighth annual international conference on Research in computational molecular biology*, pages 187–193, New York, NY, USA, 2004. ACM.

[6] C. Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, December 2005.

[7] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 91 –98, October 2009.

[8] 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.

[9] 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 *GVIZ '94 Graphics and Visualization Conference*, September 1994.

[10] 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.

[11] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[12] D. M. Dunlavy, T. M. Shead, and E. T. Stanton. Paratext: Scalable text modeling and analysis. In *Proceedings of the 19th International ACM Symposium on High Performance Distributed Computing*, pages 344–347, 2010. (34

[13] S. Geman, D. Geman, K. Abend, T. J. Harley, and L. N. Kanal. Stochastic relaxation, gibbs distributions and the bayesian restoration of images*. *Journal of Applied Statistics*, 20(5):25–62, 1993.

[14] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[15] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in semantic representation. *Psychological Review*, 114(2):211–244, April 2007.

[16] M. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, June 2003.

[17] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, August 2002.

[18] 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.

[19] J. Reisinger, A. Waters, and B. S. andRaymond J. Mooney. Spherical topic models. In *Proc. ICML*, pages 903–910. Omnipress, 2010.

[20] J. A. Wise. The ecological approach to text visualization. *Journal of the American Society of Information Science and Technology*, 50(13):1224 – 1233, 1999.

[21] 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*, volume 7243, page 72430H. SPIE, 2009.