TopicView: Visually Comparing Topic Models of Text Collections

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

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

***Abstract*—We present TopicView, an application for visually comparing and exploring multiple models of text corpora. TopicView uses multiple linked views to visually analyze both the conceptual content and the document relationships in models generated using different algorithms. To illustrate TopicView, we apply it to 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. We demonstrate the utility of TopicView's visual approach to model assessment by comparing LSA and LDA models of two example corpora.**

***Keywords*-text analysis; visual model analysis; latent semantic analysis; latent dirichlet allocation.**

I. Introduction

Latent Semantic Analysis (LSA) [1] and Latent Dirichlet Allocation (LDA) [2] are two popular mathematical approaches to modeling textual data. Questions posed by algorithm developers and data analysts working with LSA and LDA models motivated the work described in this paper: 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 algorithms contain similar document clusters? How well do document clusters found in their respective similarity graphs match human-generated clusters?

LSA and LDA models, as well as many other factor models of textual data, have much in common. They both use bag-of-words modelling, 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. Yet despite these similarities, the two algorithms generate very different models. LSA uses singular value decomposition (SVD) 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. Furthermore, 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 with negative values indicating inverse correlations. LDA produces term-topic and document-topic probability 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 and instead to focus on human consumable differences relative to how these models are used. 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 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.

In this paper we present *TopicView*, an application designed to visually compare and interactively explore LSA and LDA models from this user-based perspective. In TopicView, tabbed panels of linked views compare conceptual content (i.e. *concepts/topics*) and document relationships; both relationships between individual documents, and relationships between documents and conceptual content. In addition to describing the design and implementation of TopicView, we also present some insights on differences between LSA and LDA models gained using TopicView with two small corpora.

II. Related Work

To assess how well existing methods model human semantic memory, Griffiths et al. [7] compare generative probabilistic topic models with models of semantic spaces. 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, our work focuses on comparing the impact that model differences have on visual analytics applications, using visualization to do the comparison.

Collins et al. [8] 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. 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 feel 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. This avoids the layout complications and potential overlaps encountered when words are drawn at vastly different scales.

III. Modeling Approaches Used in this Work

A. Latent Semantic Analysis

LSA computes a truncated SVD of a term-document matrix [9], i.e., the collection of weighted term vectors associated with the documents in a corpus of text. 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.

B. Latent Dirichlet Allocation

LDA is a hierarchical probabilistic generative approach that models a collection of documents by topics, i.e., probability distributions over a vocabulary [2]. Given a vocabulary of W distinct words, a number of topics K, two smoothing parameters α and β, 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, the parameters of the generative model must be learned from the data. Specifically, for a corpus containing D documents we want to learn ϕ, the $K \times W$ matrix of topics, and θ, the $D \times K$ matrix of topic weights 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 [10] to estimate θ and ϕ.

C. 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 are stored as a similarity matrix, which 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. 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 represent clusterings of the documents, i.e., related nodes are grouped together in the resulting graph layout.

IV. TopicView

TopicView loads a corpus of documents and uses both LSA and LDA to generate models of the data. Shared preprocessing produces a term-document matrix and a term dictionary that serve as inputs to both algorithms. Identical rank (LSA) and topic counts (LDA) are input to generate matching numbers of concepts or topics in their respective outputs. 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 differences to just those attributable to the two algorithms.

Conceptual content and document relationships are visualized using separate tabbed displays. Views 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 the left document similarity graph), whereas LDA-generated model components are displayed in red and appear on the right.

A. Exploring Conceptual Content

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. 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 (i.e., correlations and probabilities), we can visually compare the ordering and relative weighting of their terms.

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 Concepts panel in Figure 1, 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).

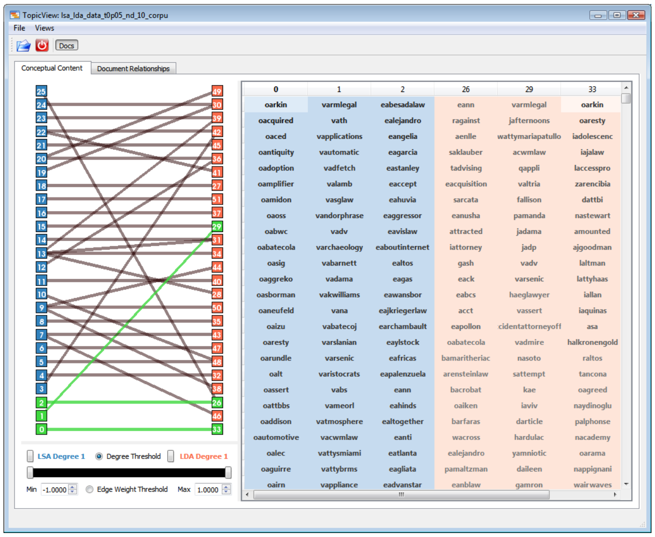
Concept/topic similarities are calculated as follows:

i) Scale LSA's left singular matrix by its singular values.
ii) Concatenate the result with LDA's θ matrix.
iii) Compute the matrix of pairwise cosine similarities of all rows of the concatenated matrix.
iv) Truncate the result 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 red (1) to black (0) to blue (-1) to preserve the distinction between positive and negative similarities.

Fixing the LSA node positions in their rank order to capture information about variance, we use a greedy approach for placing LDA nodes relative to the LSA nodes. Nodes are laid out in declining edge weight order so as to draw the strongest similarities with horizontal edges.

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



Min -1.0000 Max 1.0000

0	1	2	26	29	33
oarkin	varmlegal	eabesadalaw	eann	varmlegal	oarkin
oacquired	vath	ealejandro	ragainst	jafternoons	oaresty
oaced	vapplications	eangelia	aenlle	wattymariapatullo	iadolescenc
oantiquity	vautomatic	eagarcia	saklauber	acwmlaw	iajalaw
oadoption	vadfetch	eastanley	tadvising	qappli	laccesspro
oamplifier	valamb	eaccept	eacquisition	valtria	zarencibia
oamidon	vasglaw	eahuvia	sarcata	fallison	dattbi
oaoss	vandorphrase	eaggressor	eanusha	pamanda	nastewart
oabwc	vadv	eavislaw	attracted	jadama	amounted
oabatecola	varchaeology	eaboutinternet	iattorney	jadp	ajgoodman
oasig	vabarnett	ealtos	gash	vadv	laltman
oaggreko	vadama	eagas	eack	varsenic	lattyhaas
oasborman	vakwilliams	eawansbor	eabcs	haeglawyer	iallan
oaneufeld	vana	eajkriegerlaw	acct	vassert	iaquinas
oaizu	vabatecoj	earchambault	eapollon	:cidentattorneyoff	asa
oaresty	varslanian	eaylstock	oabatecola	vadmire	halkronengold
oarundle	varsenic	eafricas	bamaritheriac	nasoto	raltos
oalt	varistocrats	eapalenzuela	arensteinlaw	sattempt	tancona
oassert	vabs	eann	bacrobat	kae	oagreed
oattbbs	vameorl	eahinds	oaiken	iaviv	naydinoglu
oaddison	vatmosphere	ealtogether	barfaras	darticle	palphonse
oautomotive	vacwmlaw	eanti	wacross	hardulac	nacademy
oalec	vattysmiami	eatlanta	ealejandro	yamniotic	oarama
oaguirre	vattybrms	eagliata	pamaltzman	daileen	nappignani
oairn	vappliance	eadvanstar	eanblaw	gamron	wairwaves

Figure 1: 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.

Nodes and edges in the bipartite graph are selectable, with selections shown in green. Selecting an edge selects its two end nodes. Nodes correspond to columns in the *Term Table* and columns in the *Document Table* on the Document Relationships panel. As shown in both Figures 1 and 2, selection reduces the columns displayed in both tables to 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 restores the entire set of columns.

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 1. 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 can only be directly compared within the same algorithm.

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.

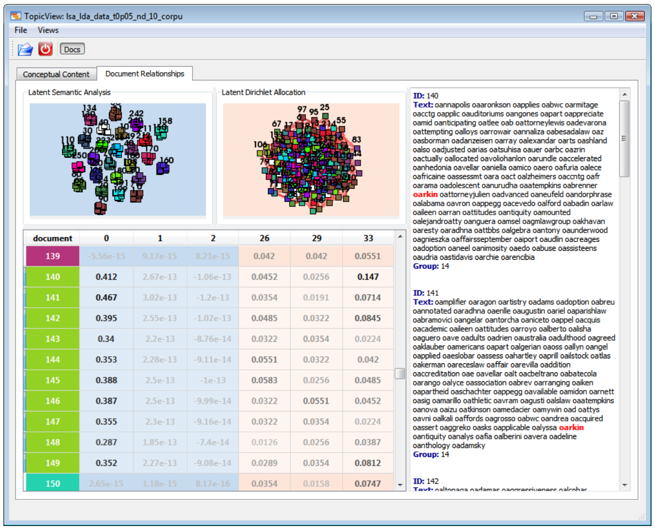
B. Viewing Document Relationships

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/topics with clusters, the weightings shown in the *Document Table* demonstrate that document clusters frequently contribute in varying degrees to multiple concepts/topics (weightings spread across rows). Similarly, concepts/topics typically include multiple document clusters (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 clusters, and then to read their full texts in the *Document Text* view.

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 weights between every document pair, so we reduce visual clutter by thresholding 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 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 2, 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 selections drawn in white. 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). The selected documents are also highlighted in the *Document Table* and their full text is displayed in the *Document Text* view.



document	0	1	2	26	29	33
139	-5.56e-15	9.17e-15	8.21e-15	0.042	0.042	0.0551
140	0.412	2.67e-13	-1.06e-13	0.0452	0.0256	0.147
141	0.467	3.02e-13	-1.2e-13	0.0354	0.0191	0.0714
142	0.395	2.55e-13	-1.02e-13	0.0485	0.0322	0.0845
143	0.34	2.2e-13	-8.76e-14	0.0322	0.0354	0.0224
144	0.353	2.28e-13	-9.11e-14	0.0551	0.0322	0.042
145	0.388	2.5e-13	-1e-13	0.0583	0.0256	0.0485
146	0.387	2.5e-13	-9.99e-14	0.0322	0.0551	0.0452
147	0.355	2.3e-13	-9.16e-14	0.0322	0.0354	0.0224
148	0.287	1.85e-13	-7.4e-14	0.0126	0.0256	0.0387
149	0.352	2.27e-13	-9.08e-14	0.0289	0.0354	0.0812
150	2.65e-15	1.18e-15	8.17e-16	0.0354	0.0158	0.0747

ID: 140
Text: oannapolis oaaronkson oapplies oabwc oarmitage oacctg oapplic oauditoriums oangones oapart oappreciate oamid oanticipating oatlee oab oattorneylewis oadevarona oattempting oalloys oarrowair oannaliza oabesadalaw oaz oasborman oadanzeisen oarray oalexandar oarts oashland oalso oadjusted oarias oatsuhisa oauer oarbc oazrin oactually oallocated oavoliohanlon oarundle oaccelerated oanhedonia oavellar oaniella oamico oaero oafuria oalece oafricaine oassessmt oara oact oalzheimers oaccntg oafr oarama oadolescent oanurudha oaatempkins oabrenner oarkin oattorneyjulien oadvanced oaneufeld oandorphrase oalabama oavron oappegg oacevedo oalford oabadin oarlaw oaileen oarran oattitudes oantiquity oamounted oalejandroatty oanguera oamsel oagmlawgroup oakhavan oaresty oaradhna oattbbs oalgebra oantony oaunderwood oagnieszka oaffairsseptember oaiport oaudlin oacreages oadoption oaneel oanimosity oaedo oabuse oassisteens oaudria oastidavis oarchie oarencibia
Group: 14

ID: 141
Text: oamplifier oaragon oartistry oadams oadoption oabreu oannotated oaradhna oaenlle oaugustin oariel oaparishlaw oabramovici oangelar oantorcha oaniceto oappel oacquis oacademic oaileen oattitudes oarroyo oalberto oalisha oaguero oave oadults oadrien oaustralia oadulthood oagreed oaklauber oamericans oapart oalgerian oaoss oallyn oangel oapplied oaeslobar oassess oahartley oaprill oailstock oatlas oakerman oareceslaw oaffair oarevilla oaddition oaccreditation oae oavellar oalt oacbeltrano oabatecola oarango oalyce oassociation oabrev oarranging oaiken oapartheid oaschachter oappegg oavailable oamidon oarnett oasig oamarillo oathletic oavram oagusti oalslaw oaatempkins oanova oaizu oatkinson oamedacier oamywin oad oattys oavni oalkali oaffords oagrosso oabwc oandrea oacquired oassert oaggreko oasks oapplicable oalyssa oarkin oantiquity oanalys oafia oalberini oavera oadeline oanthology oadamsky
Group: 14

ID: 142
Text: oaltonaga oadamas oaggressiveness oalcobar

Figure 2: 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.

2) Document Table: The *Document Table* (lower left Figure 2), shows the concatenation of the transpose of LSA’s right singular vector, scaled by its singular values, with 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 both graphs and display the selected document contents in the *Document Text* view.

3) Document Text: Provides the full text contents of multiple documents, selected using the *Document Similarity Graphs* or the *Document Table*. Each document is displayed as three fields: the document ID, the raw text of the document, 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.

V. 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 corpora. The goals of the case studies include the following:

- Illustrate the use of TopicView for efficient 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.

The LSA and LDA models were computed using the ParaText library [11] from the open source Titan Informatics Toolkit [12]. For LDA, we set $\alpha = 50 / K$, $\beta = 0.1$, sampling iterations = 1, and burn in iterations = 200.

Two data sets were used in the case studies. The first case study used the artificial *alphabet* data set, in which the terms in each cluster are entirely disjoint from one another. The second case study used the *DUC* data set, a real-world document collection in which terms and concept/topics overlap normally. Both data sets have human-generated cluster labels, which are used to color-code the document groups to identify how well the algorithms’ clusters match human-generated ground truth.

The *alphabet* data set consists of 26 clusters containing 10 documents each. Each cluster consists of documents made up exclusively of terms starting with the same letter. 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 consisting of 26000 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 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 [13]. The collection consists of 298 documents categorized into 30 clusters, with each cluster containing roughly 10 documents focused on a particular topic or event.

A. *Case Study Using Alphabet Data*

As LSA concepts are, by definition, orthogonal latent feature vectors, we expect that LSA should be able to model each of the document clusters in the *alphabet* data set 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 cluster the data.

Figure 3 presents TopicView's *Document Similarity Graphs* for the LSA (left) and LDA (right) models applied to the *alphabet* data set. We see that the LSA model clusters the *alphabet* data set well; there are 26 disconnected components in the graph, and each 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. As shown in Figure 1, the *Term Table* on the Conceptual Content panel 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). In the *Term Table*, we can see that words beginning with "o", "v", and "e" are 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 panel confirm that LSA models the clusters correctly; see Figure 2 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 LSA concepts and LDA topics. Figure 4 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., rows of θ), 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

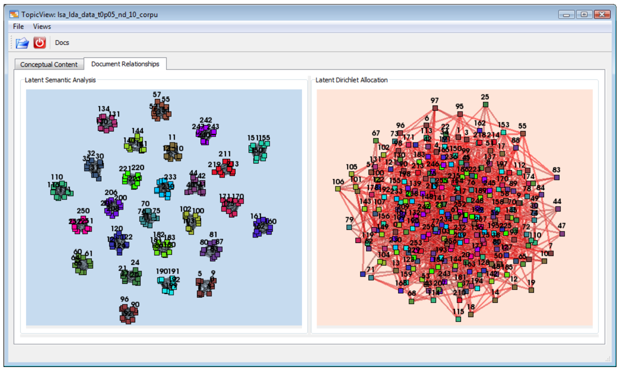


Figure 3: Graphs depicting document relationships modeled using LSA (left) and LDA (right) for the *alphabet* data set.

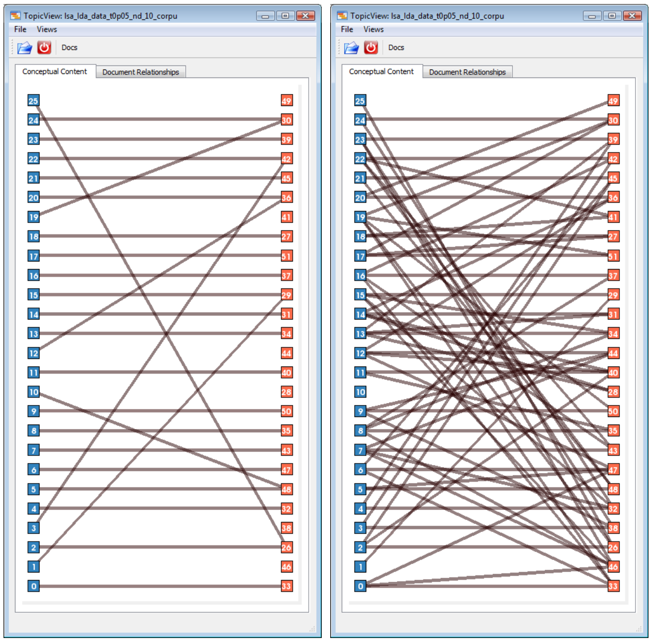


Figure 4: Graphs depicting conceptual contents relationships for the *alphabet* data set modeled with LSA (blue nodes) and LDA (red nodes) using a minimum degree threshold of 1 for LSA concept nodes (left) and a minimum edge threshold of 0.1451 (right).

the relationships between 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 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 not at all related to the LSA concepts, each of which models one of the 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 capturing the term 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, the topics from LDA tend to be random with more or less uniform term distributions.

B. Case Study Using DUC Data

The case study in the previous section 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 do not appear often in real-world analysis applications. Even when corpora 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 documents whose general topics appeared 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 different in how weak connections between document clusters are modeled.

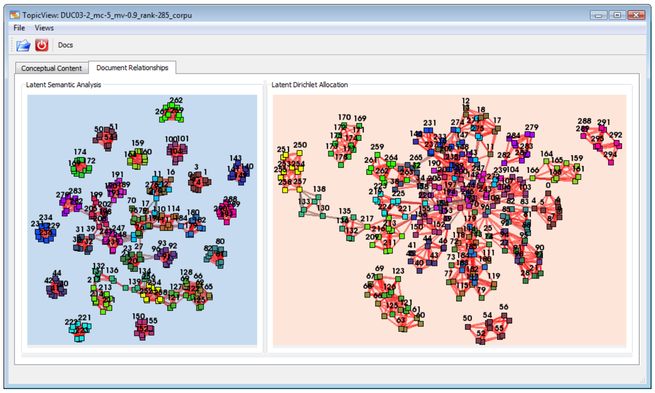
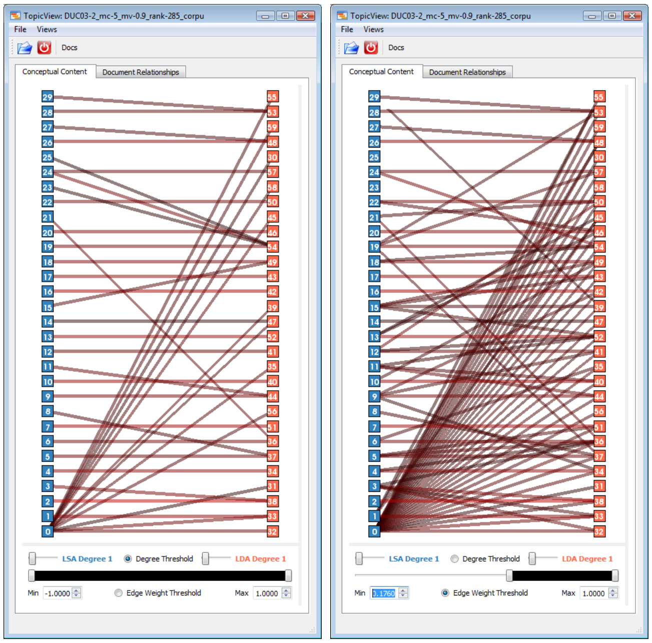


Figure 5: Graphs depicting document relationships modeled using LSA (left) and LDA (right) for the *DUC* data set.



● Degree Threshold, LSA Degree 1, LDA Degree 1, Min -1.0000, ○ Edge Weight Threshold, Max 1.0000

○ Degree Threshold, LSA Degree 1, LDA Degree 1, Min 0.1760, ● Edge Weight Threshold, Max 1.0000

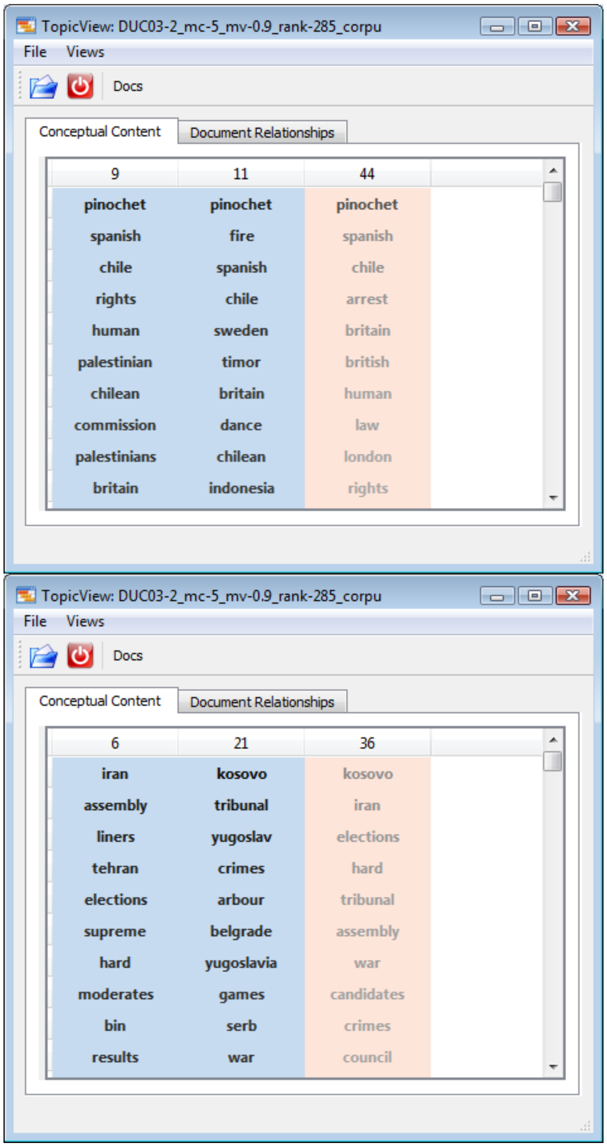
Figure 6: Graphs depicting conceptual content relationships for the *DUC* data set modeled with LSA (blue nodes) and LDA (red nodes) using a minimum degree threshold of 1 for LSA/LDA nodes (left) and a minimum edge threshold of 0.1760 (right).

As in the previous case study, we start our analysis of the relationships between the LSA and LDA models by examining the document similarity graphs for the two models (Figure 5). The LSA model (left graphs) results in a graph with 14 disconnected components, 13 of which match 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. Note 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 (Figure 5), 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 6 illustrates how TopicView's Conceptual Content Bipartite Graph views can be used to indicate 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 is modeling the statistical variance of terms across the documents and thus LSA concept 0 acts as a generic concept that summarizes all of the main interactions between documents as a function of the terms appearing in those documents [1].

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. 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 7, 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 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 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 (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 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.



9	11	44
pinochet	pinochet	pinochet
spanish	fire	spanish
chile	spanish	chile
rights	chile	arrest
human	sweden	britain
palestinian	timor	british
chilean	britain	human
commission	dance	law
palestinians	chilean	london
britain	indonesia	rights

6	21	36
iran	kosovo	kosovo
assembly	tribunal	iran
liners	yugoslav	elections
tehran	crimes	hard
elections	arbour	tribunal
supreme	belgrade	assembly
hard	yugoslavia	war
moderates	games	candidates
bin	serb	crimes
results	war	council

Figure 7: Top 10 terms associated with the concepts/topics where multiple LSA concepts are strongly connected to a single LDA topic.

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 this case that LSA and LDA model the most tightly coupled document clusters well (as indicated by many strong horizontal edges in Figure 6), but model more subtle relationships between documents (and thus clusters) in different ways. Mathematically, we conjecture that this is because LSA's singular vectors involve both positive and negative weights for a term. This lets components of one concept "cancel out" components of another. By comparison, the term weights in LDA topics are strictly positive: 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. 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 model differences.

VI. Conclusions and Future Work

Using TopicView, we find 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 more or less vocabulary overlap, to investigate the generality of the findings presented in the two case studies 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) [14] and extensions to LDA that have shown improved performance in document clustering applications, such as mixture of von Mises-Fisher models [15].

Acknowledgments

This work was funded 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*, vol. 41, no. 6, pp. 391–407, 1990.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 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*, vol. 57, no. 3, pp. 359–377, December 2005.

[5] J. A. Wise, "The ecological approach to text visualization," *JASIST*, vol. 50, no. 13, pp. 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*, vol. 11, no. 3, pp. 259–285, 1998.

[7] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation," *Psychological Review*, vol. 114, no. 2, pp. 211–244, April 2007.

[8] C. Collins, F. B. Viégas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *Proc. VAST*, October 2009, pp. 91 –98.

[9] M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Review*, vol. 37, no. 4, pp. 573–595, 1995.

[10] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *PNAS*, vol. 101, no. Suppl. 1, pp. 5228–5235, April 2004.

[11] D. M. Dunlavy, T. M. Shead, and E. T. Stanton, "Paratext: Scalable text modeling and analysis," in *Proc. HPDC*, 2010, pp. 344–347.

[12] B. Wylie and J. Baumes, "A unified toolkit for information and scientific visualization," in *Proc. Visualization and Data Analysis*, vol. 7243. SPIE, 2009, p. 72430H.

[13] 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.

[14] M. Berry and M. Browne, "Email surveillance using nonnegative matrix factorization," *Comput. Math. Organ. Th.*, vol. 11, no. 3, pp. 249–264, Oct. 2005.

[15] J. Reisinger, A. Waters, B. Silverthorn, and R. J. Mooney, "Spherical topic models," in *Proc. ICML*, 2010, pp. 903–910.