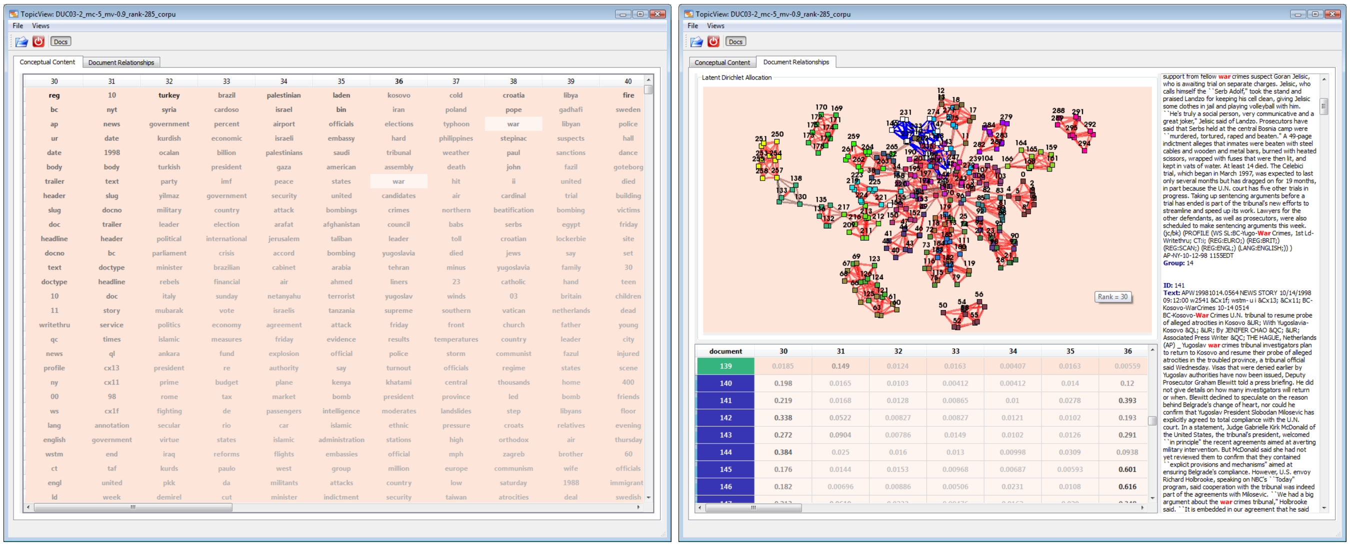
TopicView: Understanding Document Relationships Using Latent Dirichlet Allocation Models

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


TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu
File
Views
Docs
Conceptual Content
Document Relationships
TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu
File
Views
Docs
Conceptual Content
Document Relationships
Latent Dirichlet Allocation
Rank = 30

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

document	30	31	32	33	34	35	36
139	0.0185	0.149	0.0124	0.0163	0.00407	0.0163	0.00559
140	0.198	0.0165	0.0103	0.00412	0.00412	0.014	0.12
141	0.219	0.0168	0.0128	0.00865	0.01	0.0278	0.393
142	0.338	0.0522	0.00827	0.00827	0.0121	0.0102	0.193
143	0.272	0.0904	0.00786	0.0149	0.0102	0.0126	0.291
144	0.384	0.025	0.016	0.013	0.00998	0.0309	0.0938
145	0.176	0.0144	0.0153	0.00968	0.00687	0.00593	0.601
146	0.182	0.00696	0.00886	0.00506	0.0231	0.0108	0.616

support from fellow war crimes suspect Goran Jelisic, who is awaiting trial on separate charges. Jelisic, who calls himself the ``Serb Adolf,'' took the stand and praised Landzo for keeping his cell clean, giving Jelisic some clothes in jail and playing volleyball with him. ``He's truly a social person, very communicative and a great joker,'' Jelisic said of Landzo. Prosecutors have said that Serbs held at the central Bosnia camp were ``murdered, tortured, raped and beaten.'' A 49-page indictment alleges that inmates were beaten with steel cables and wooden and metal bars, burned with heated scissors, wrapped with fuses that were then lit, and kept in vats of water. At least 14 died. The Celebici trial, which began in March 1997, was expected to last only several months but has dragged on for 19 months, in part because the U.N. court has five other trials in progress. Taking up sentencing arguments before a trial has ended is part of the tribunal's new efforts to streamline and speed up its work. Lawyers for the other defendants, as well as prosecutors, were also scheduled to make sentencing arguments this week. (jc/bk) (PROFILE (WS SL:BC-Yugo-War Crimes, 1st Ld-Writethru; CT:i; (REG:EURO;) (REG:BRIT;) (REG:SCAN;) (REG:ENGL;) (LANG:ENGLISH;)))
AP-NY-10-12-98 1155EDT
Group: 14

ID: 141
Text: APW19981014.0564 NEWS STORY 10/14/1998 09:12:00 w2541 &Cx1f; wstm- u i &Cx13; &Cx11; BC-Kosovo-WarCrimes 10-14 0514
BC-Kosovo-War Crimes U.N. tribunal to resume probe of alleged atrocities in Kosovo &UR; With Yugoslavia-Kosovo &QL; &UR; By JENIFER CHAO &QC; &UR; Associated Press Writer &QC; THE HAGUE, Netherlands (AP) _ Yugoslav war crimes tribunal investigators plan to return to Kosovo and resume their probe of alleged atrocities in the troubled province, a tribunal official said Wednesday. Visas that were denied earlier by Yugoslav authorities have now been issued, Deputy Prosecutor Graham Blewitt told a press briefing. He did not give details on how many investigators will return or when. Blewitt declined to speculate on the reason behind Belgrade's change of heart, nor could he confirm that Yugoslav President Slobodan Milosevic has explicitly agreed to total compliance with the U.N. court. In a statement, Judge Gabrielle Kirk McDonald of the United States, the tribunal's president, welcomed ``in principle'' the recent agreements aimed at averting military intervention. But McDonald said she had not yet reviewed them to confirm that they contained ``explicit provisions and mechanisms'' aimed at ensuring Belgrade's compliance. However, U.S. envoy Richard Holbrooke, speaking on NBC's ``Today'' program, said cooperation with the tribunal was indeed part of the agreements with Milosevic. ``We had a big argument about the war crimes tribunal,'' Holbrooke said. ``It is embedded in our agreement that he said

Fig. 1: The TopicView user interface. At left, the Conceptual Content panel presents a Term Table with topics as columns, listing terms in decreasing order of importance within each topic. The Document Relationship panel on the right displays the Document Similarity Graph at the top, the Document-Topic Table below, and full text for selected documents on the right. Words selected in the Topic-Term view are highlighted in red in the document text. Near the top of the Document Similarity Graph, articles on Kosovan war crimes and Iranian elections are highlighted in white and the edges linking them are highlighted in blue.

Abstract—Document similarity graphs are a useful visual metaphor for assessing the conceptual content of a corpus. Algorithms such as Latent Dirichlet Allocation (LDA) provide a means for constructing such graphs by extracting topics and their associated term lists, which can be converted into similarity measures. Given that users' understanding of the corpus content (and therefore their decision-making) depends upon the outputs provided by LDA as well as how those outputs are translated into a visual representation, an examination of how the LDA algorithm behaves and an understanding of the impact of this behavior on the final visualization is critical. We examine some puzzling relationships between documents with seemingly disparate topics that are linked in LDA graphs. We use *TopicView*, a visual analytics tool, to uncover the source of these unexpected connections.

Index Terms—Visual analytics, Text analysis, Latent Dirichlet Allocation.

1 Introduction

In working with similarity graphs generated from Latent Dirichlet Allocation (LDA) [1] models, we were struck by the high degree of connectivity in the graph. Many of the edges were unexpected, connecting documents covering seemingly unrelated topics. Why was LDA linking these documents? We felt that understanding the cause of these links was essential, since their presence alters both the graph layout and its interpretation by users [5].

To explore this problem, we used visual analytics to reveal how the LDA algorithm makes the connections shown in a similarity graph. Our tool *TopicView* [2] combines a user-level view of the similarity graph with linked views that enable exploration of the relationships between documents, topics, and terms. This paper describes our investigation's visual analysis and the improved separation of topic groups we were able to achieve as a result.

2 Background Overview

This section provides a brief description of the data set that we used in our analysis, the LDA algorithm and associated parameter values, and the *TopicView* visualizations used.

2.1 Data Set

For this analysis, we used the *DUC* data set, a collection of newswire documents from the Associated Press and New York Times that were used in the 2003 Document Understanding Conference (DUC) to evaluate document summarization systems [4]. The collection contains 30 clusters comprising 298 documents. Each cluster contains roughly 10 documents focused on a particular topic or event. The documents all have human-generated cluster labels, which we use to color-code documents. We informally define a cluster as a group of documents exhibiting strong links between members within the group and weak links outside the group. As with any non-trivial, real-world corpus, the topics in each cluster are not entirely disjoint in term distributions; rather, each topic focuses upon a separate news event. Thus, there are some conceptual commonalities that reasonably can be considered as possible linking mechanisms between clusters.

- *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.

2.2 Latent Dirichlet Allocation

LDA is a hierarchical probabilistic generative model used to model a collection of documents by topics, i.e., probability distributions over a vocabulary [1]. Given a vocabulary of W distinct words, a number of topics K, two smoothing parameters α and β, and a prior distribution (typically Poisson) over document lengths, 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 [3] to estimate θ and ϕ.

We set $K = 30$ to match the number of anticipated clusters in the corpus. Following Blei et al. [1], we use $\alpha = 50/K$ and $\beta = 0.1$. Two additional parameters for the Gibbs sampling are the number of sampling and burn-in iterations, which we set to 1 and 200, respectively.

2.3 TopicView

Although TopicView was developed to compare LSA models with LDA models [2], for this paper we use it to explore individual LDA models. We predominently use the following four views: the *Term Table*, the *Document Similarity Graph*, the *Document-Topic Table*, and *Document Text*. Documents can be selected either through the graph of document relationships or the table of document model features, highlighting the selection in both views and displaying the selected document contents in the *Document Text* view.

2.3.1 Term Table

The *Term Table* shown in Figure 2 presents the terms for each topic (i.e., the rows of the topic matrix ϕ) sorted in decreasing order of importance. Text color provides an additional cue about the relative weights of terms. Terms with the highest weights are drawn in black, fading to gray for the lowest-weighted terms. 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 logarithmic mapping that increases the number of luminance steps as we approach black. Individual terms are selectable. Once selected, each instance of that term within every 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 highlighted in red.

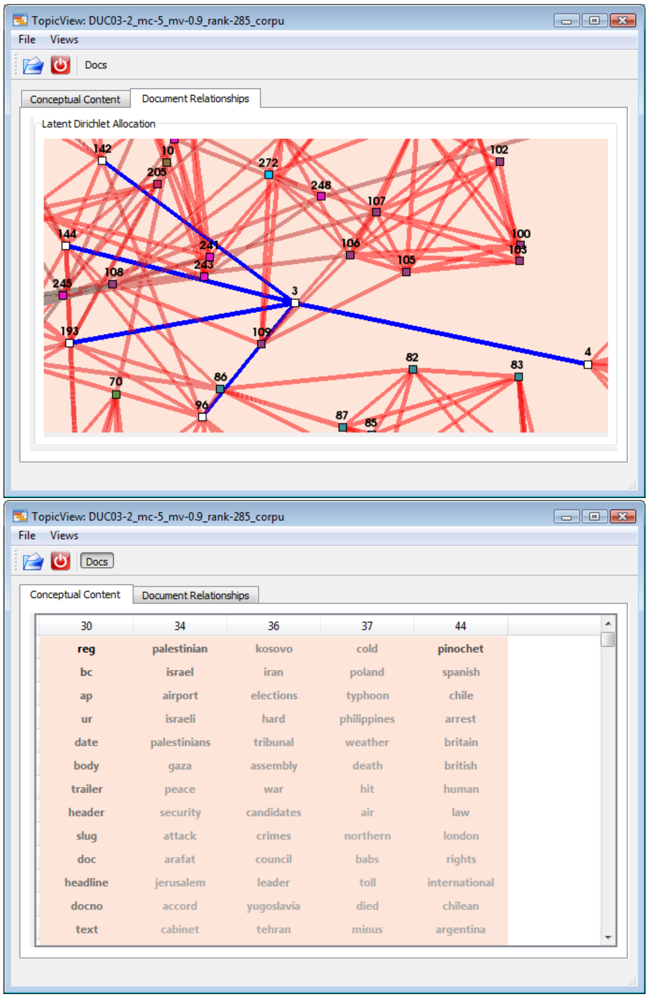
2.3.2 Document Similarity Graphs

In TopicView, we compute edge weights between every pair of documents by calculating the cosine similarities of the topic weight matrix θ. To reduce visual clutter, we threshold edges by keeping the strongest links, while retaining a minimum number of edges per node. We determine which edges to keep on a document-by-document basis as follows: (i) sort each document's edges in descending order by weight, (ii) keep all edges with weights greater than a significance threshold (we use 0.9), and (iii) if a document's edge count is less than a minimum (5 in our examples), add edges in diminishing weight order until it is reached.

We project the graphs into two dimensions using a linear time force-directed layout algorithm. Each document node is labeled with its document ID and color-coded using the ground-truth category from *DUC*. Edges are color-coded using saturation to indicate similarity weights, with low values in gray and high values in red. We highlight selected nodes in white, which is the one color not used for the DUC label categories. Selected edges are in blue.

2.3.3 Document-Topic Table

The *Document-Topic Table* shown in Figure 3 is LDA's θ matrix of topic weights for each document. 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 topic. Although there is a tendency to try to identify topics with clusters, the weightings shown in the *Document Table* demonstrate that document groups frequently contribute in varying degrees to multiple topics (weightings spread across rows). Similarly, topics typically include multiple document groups (weightings spread across columns).



30	34	36	37	44
reg	palestinian	kosovo	cold	pinochet
bc	israel	iran	poland	spanish
ap	airport	elections	typhoon	chile
ur	israeli	hard	philippines	arrest
date	palestinians	tribunal	weather	britain
body	gaza	assembly	death	british
trailer	peace	war	hit	human
header	security	candidates	air	law
slug	attack	crimes	northern	london
doc	arafat	council	babs	rights
headline	jerusalem	leader	toll	international
docno	accord	yugoslavia	died	chilean
text	cabinet	tehran	minus	argentina

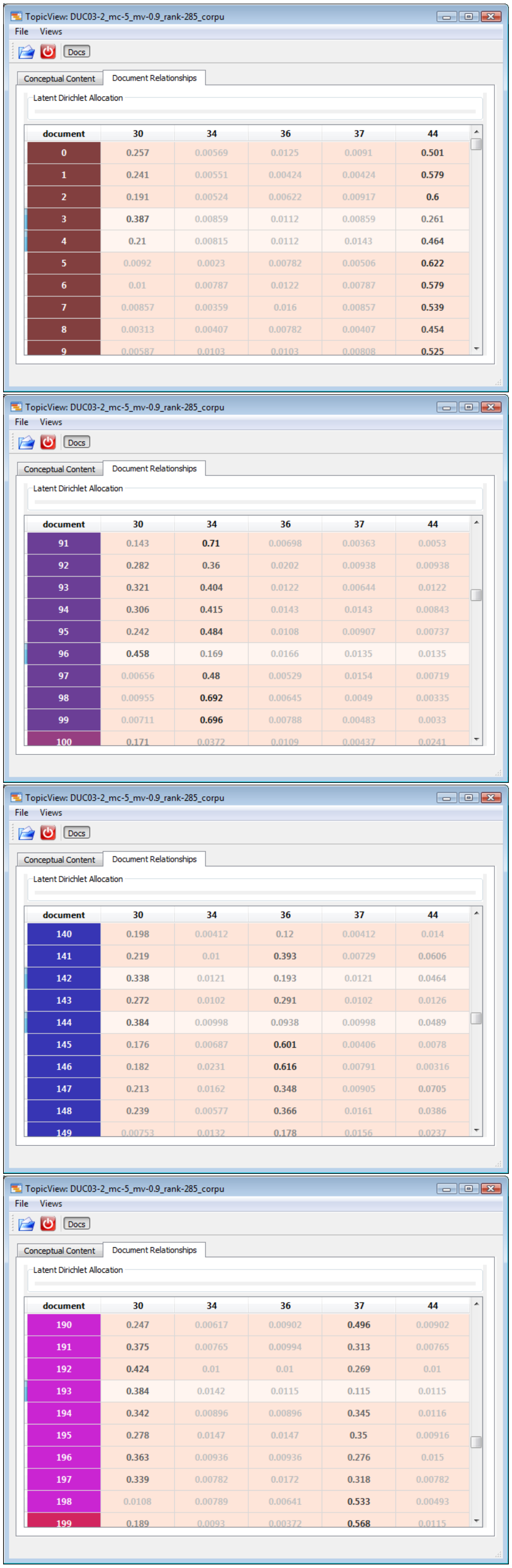
Fig. 2: Top image: The Document Similarity Graph displays document 3 and the connections to its immediate neighbors highlighted in white (nodes) and blue (edges). Bottom image: The Term Table displays topics associated with the selected documents shown in the Document-Topic Table images in Figure 3. The topics are chosen based on the strength of the document-topic weights for the selected documents.

The visible columns within the table are controlled by selecting topics. Subsetting the column display facilitates side-by-side comparisons of the relative weightings of the most significant documents associated with a set of topics. The table can be used to formulate hypotheses about the relationship between conceptual content and specific documents.

3 Analysis

We start by examining documents with seemingly unrelated topics, such as document 3 shown in the center of the graph in Figure 2. According to the text of documents 3 and 4, both are stories about Pinochet's arrest in Britain. Document 96 is a story about Israel delaying flights from a Palestinian airport. Documents 142 and 144 are about a Yugoslavian tribunal to prosecute Bosnian war crimes. Finally, document 193 is about cold weather killing 39 people in Moscow. Identifying the connection between these documents will provide insight into document characteristics modeled by LDA.

The stories in the selected documents are represented by topics 44, 34, 36, and 37, respectively. Terms capturing these key concepts can be seen amongst the top terms for these topics in Figure 2. (We will


TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu
File
Views
Docs
Conceptual Content
Document Relationships
Latent Dirichlet Allocation

document	30	34	36	37	44
0	0.257	0.00569	0.0125	0.0091	0.501
1	0.241	0.00551	0.00424	0.00424	0.579
2	0.191	0.00524	0.00622	0.00917	0.6
3	0.387	0.00859	0.0112	0.00859	0.261
4	0.21	0.00815	0.0112	0.0143	0.464
5	0.0092	0.0023	0.00782	0.00506	0.622
6	0.01	0.00787	0.0122	0.00787	0.579
7	0.00857	0.00359	0.016	0.00857	0.539
8	0.00313	0.00407	0.00782	0.00407	0.454
9	0.00587	0.0103	0.0103	0.00808	0.525

document	30	34	36	37	44
91	0.143	0.71	0.00698	0.00363	0.0053
92	0.282	0.36	0.0202	0.00938	0.00938
93	0.321	0.404	0.0122	0.00644	0.0122
94	0.306	0.415	0.0143	0.0143	0.00843
95	0.242	0.484	0.0108	0.00907	0.00737
96	0.458	0.169	0.0166	0.0135	0.0135
97	0.00656	0.48	0.00529	0.0154	0.00719
98	0.00955	0.692	0.00645	0.0049	0.00335
99	0.00711	0.696	0.00788	0.00483	0.0033
100	0.171	0.0372	0.0109	0.00437	0.0241

document	30	34	36	37	44
140	0.198	0.00412	0.12	0.00412	0.014
141	0.219	0.01	0.393	0.00729	0.0606
142	0.338	0.0121	0.193	0.0121	0.0464
143	0.272	0.0102	0.291	0.0102	0.0126
144	0.384	0.00998	0.0938	0.00998	0.0489
145	0.176	0.00687	0.601	0.00406	0.0078
146	0.182	0.0231	0.616	0.00791	0.00316
147	0.213	0.0162	0.348	0.00905	0.0705
148	0.239	0.00577	0.366	0.0161	0.0386
149	0.00753	0.0132	0.178	0.0156	0.0237

document	30	34	36	37	44
190	0.247	0.00617	0.00902	0.496	0.00902
191	0.375	0.00765	0.00994	0.313	0.00765
192	0.424	0.01	0.01	0.269	0.01
193	0.384	0.0142	0.0115	0.115	0.0115
194	0.342	0.00896	0.00896	0.345	0.0116
195	0.278	0.0147	0.0147	0.35	0.00916
196	0.363	0.00936	0.00936	0.276	0.015
197	0.339	0.00782	0.0172	0.318	0.00782
198	0.0108	0.00789	0.00641	0.533	0.00493
199	0.189	0.0093	0.00372	0.568	0.0115

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

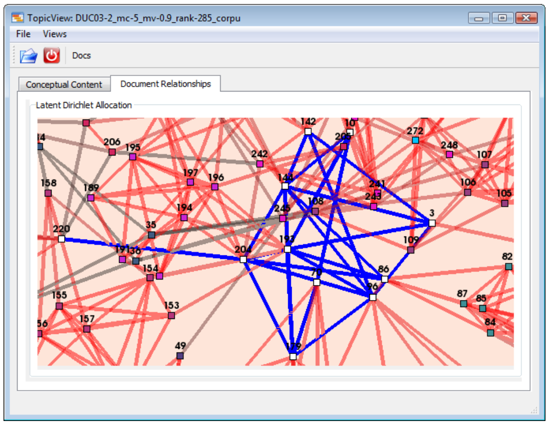


Fig. 4: All of the bridging documents are in white with the connecting edges in blue.

discuss additional terms such as Iran in topic 36 later.) The selection of these four topics comes from an examination of the document-topic weights for each document shown in Figure 3. We have included all topic columns that have darker/higher weights within the selected rows, including topic 30, whose most significant terms do not match any of the story lines seen in the documents' texts. Examining the terms in topic 30, we find that XML document tags ("headline", "slug", "body", etc.) predominate. As shown in the left image in Figure 1, topics 30 and 31 capture Associated Press (AP) and New York Times (NYT) document origins, respectively. The connection between these seemingly dissimilar documents is that they are all AP articles.

3.1 Bridging Documents

Looking at the weights in the Document-Topic Table images in Figure 3, an interesting pattern emerges. Documents 0–9 (top image) are articles about Pinochet's arrest. The human-generated cluster labels (brown color-coding in the document ID column) show that these documents all belong to this group. LDA has similarly identified this same group of documents as a group, shown by the darker text of the stronger weights in column 44. The weights for documents 0 through 4 in column 30 show a significant connection between these documents and the topic, articles from AP, whereas documents 5 to 9 do not. Unlike document 4, the weighting for document 3 in topic 30 is stronger than its weighting in topic 44 (i.e. document 3 is more strongly aligned with its AP source than with its conceptual content). This same pattern is seen with documents 96, 142, 144, and 193. We hypothesize that documents matching this pattern are the source of many of our edges between disjoint topics.

We test our hypothesis by listing all documents whose strongest weights are within the AP or NYT topic columns (30 and 31), then we check the conceptual content of these documents against all documents directly linked to them in the Document Similarity Graph. If the linked documents can be seen as have a connection in terms of their content, then the list document fails the hypothesis and is removed. We define a conceptual connection to exist between a list and linked document if both are strongly weighted in the same topic column in the Document-Topic Table, if a number of common terms are found in their document texts, or if their human-labeled categories match.

Of the original 297 documents, 33 fit our hypothesis. Of those, only 11 survived our test and exhibited links that we could not account for in some other way, including the ones originally observed (3, 96, 142, 144, and 193). The 11 documents and the edges connecting them are shown in Figure 4. These documents are in the center of the graph and all of them tend to link with one other, impacting the layout of their associated clusters and creating so many edge crossings that the true connectivity is difficult to follow. The full graph is shown in Figure 1.

3.2 Tag Terms

All of the bridging documents are AP articles. All are short, with the story content sometimes being only a single sentence. Comparing

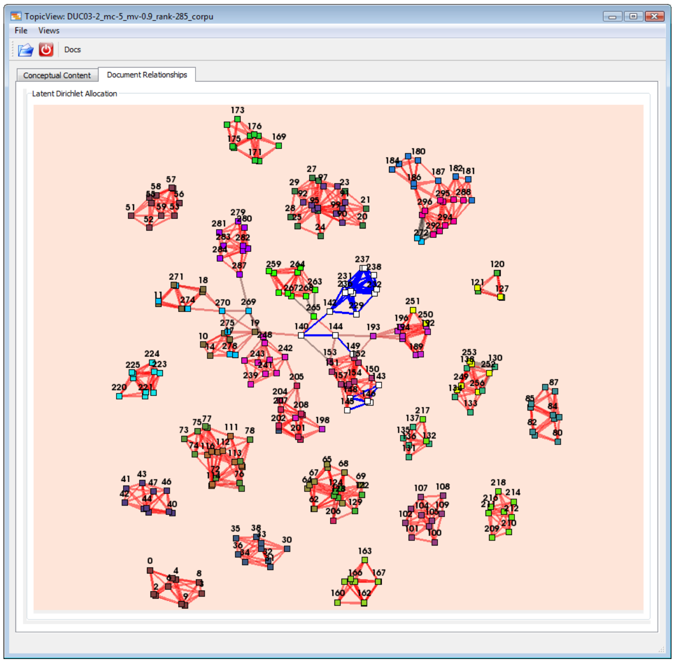

TopicView: DUC03-2_mc-5_mv-0.9_rank-285_corpu
File
Views
Docs
Conceptual Content
Document Relationships
Latent Dirichlet Allocation

Fig. 5: Document Similarity Graph for LDA run on headers and story bodies only. Articles on Kosovan war crimes and Iranian elections are highlighted in white with blue edges.

AP articles to NYT articles, we find that the Times articles tend to be longer, sometimes much longer. Both sets of articles contain tags and header information, but for these short articles, the terms in the headers outnumber the news content. Document 96 is a good example of this, being very short and highly weighted in topic 30. Thus we revise our hypothesis: only documents whose conceptual content is outweighed by the source content will display this bridging pattern.

To test the revised hypothesis, we reran LDA on just the text from the headlines and story bodies. As expected, the AP and NYT topics disappeared, many of the edges connecting unrelated topics disappeared and clusters became stronger and more visible in the graph, as shown in Figure 5.

3.3 Merged Clusters and Bridge Clusters

Now we return to the combination of articles about Kosovan war crimes (document IDs in the 140s) and Iranian elections (230s) in topic 36. This combined topic generates edges between all of the documents in both clusters. In Figure 1, these documents and their linking edges are highlighted in white in the Document Similarity Graph on the right, and the terms for topic 36 are shown in the Term Table on the left. Given that two topics are taken up by AP and NYT, the number of topics in the analysis that included headers and tags should be insufficient to handle the expected number of clusters. In the subsequent run, two new topics are available and these clusters are separated into different topics, 34 and 40 (note that the new topic numbering has no relationship to the old). The new graph is shown in Figure 5, where both sets of documents and their connections are highlighted in white.

Now only document 142 joins the two clusters. Examining the document-topic weightings reveals that 142 is more strongly linked to the Iranian election topic, 40, than the Kosovo tribunal topic. Document 142 is short (it was one of the 11 linking documents above), so it has relatively few terms to define it. Examining the top terms in topic 40, we find that *assembly* and *candidates* are the second and third most highly ranked terms, and they appear in document 142 three and four times each. The story is about a UN General Assembly vote to select judges for the war crimes tribunal. It discusses the candidates and is really an election story, which validates its connection to the Iranian election articles. So in this case, the remaining edges are legitimate.

However, not all of the clusters seen in the new graph agree with the human-labeled cluster boundaries. Once again, seemingly unrelated stories are combined (Chechen kidnappings are added to Kosovo war crimes in topic 34), or a small number of documents sharing some common aspect from multiple clusters are combined to form a topic with a much narrower focus. The latter situation connects a number of clusters through topic 58, which links the trial-related stories in documents 19, 10, 144, and 265 (found near the center of the graph in Figure 5). Here, in a manner similar to the AP topic, the individual documents are more strongly connected to each other through this subtopic than they are to their cluster groups. This topic acts as a bridge connecting all of the clusters together.

Clearly, our initial choice of 30 topics is impacting the resulting clusters, but given LDA's merging and splitting of topics, it is difficult to select an appropriate value. Experimenting with various topic values between 28 and 75, we find that increasing the topic count does not necessarily separate combined clusters or reduce the number of edges. We find new bridging topics, and topics that combine into new merged clusters. In addition, as the number of topics expands, the document weightings for a subset of the additional topics becomes so low as to make the topics appear to be noise.

4 Conclusions

LDA's choice of topics may include unexpected categories, such as article sources, if the document text contains term distributions that can act as signatures for those sources. For example, using LDA to model VisWeek's digital proceedings generates a topic consisting of all the HTTP references. In turn, these additional topics act as bridges between conceptual topics, linking seemingly unrelated articles. Short documents facilitate this by being more strongly connected to their source topic than to their conceptual content. Although we were able to remove source-specific terms from our documents and generate only thematic topics, this only works if source-based and concept-based terminology differ. If the various sources have unique writing styles that rely on a distinct vocabulary, this filtering may not be possible. The issues around selecting the best number of topics and counteracting the tendency of short documents to act as bridges between dissimilar document groups remain to be solved in future work.

Whether this bridging is seen as having a positive or negative impact depends upon the application. If the user is trying to understand just the thematic content of a corpus, additional document connections blur the thematic boundaries and, in the worst case, obscure the very patterns the user is hoping to see. However, if the user is trying to connect the dots between disparate bits of information, where sources provide important clues or impact the reliability of the answer, source connections may be desirable. Understanding these subtleties allows application designers to make conscious choices about the combined impact of the analysis and the visual representation on the users' understanding of the data.

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

[2] P. J. Crossno, A. T. Wilson, T. M. Shead, and D. M. Dunlavy. Topicview: Visually comparing topic models of text collections. In *Proceedings 23rd IEEE International Conference on Tools with Artificial Intelligence*, 2011.

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

[4] 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.

[5] C. Ziemkiewicz and R. Kosara. Laws of attraction: From perceptual forces to conceptual similarity. *IEEE Transactions on Visualization and Computer Graphics*, 16:1009–1016, November 2010.