Concurrent Visualization of Relationships between Words and Topics in Topic Models

Alison Smith[*], **Jason Chuang**[†], **Yuening Hu**[*], **Jordan Boyd-Graber**[*], **Leah Findlater**[*]

[*]University of Maryland, College Park, MD
[†]University of Washington, Seattle, WA

amsmit@cs.umd.edu, jcchuang@cs.washington.edu, ynhu@cs.umd.edu, jbg@umiacs.umd.edu, leahkf@umd.edu

Abstract

Analysis tools based on topic models are often used as a means to explore large amounts of unstructured data. Users often reason about the correctness of a model using relationships between words within the topics or topics within the model. We compute this useful contextual information as term co-occurrence and topic co-variance and overlay it on top of standard topic model output via an intuitive interactive visualization. This is a work in progress with the end goal to combine the visual representation with interactions and online learning, so the users can directly explore (a) why a model may not align with their intuition and (b) modify the model as needed.

1 Introduction

Topic modeling is a popular technique for analyzing large text corpora. A user is unlikely to have the time required to understand and exploit the raw results of topic modeling for analysis of a corpus. Therefore, an interesting and intuitive visualization is required for a topic model to provide added value. A common topic modeling technique is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which is an unsupervised algorithm for performing statistical topic modeling that uses a "bag of words" approach. The resulting topic model represents the corpus as an unrelated set of topics where each topic is a probability distribution over words. Experienced users who have worked with a text corpus for an extended period of time often think of the thematic relationships in the corpus in terms of higher-level statistics such as (a) inter-topic correlations or (b) word correlations. However, standard topic models do not explicitly provide such contextual information to the users.

Existing tools based on topic models, such as Topical Guide (Gardner et al., 2010), TopicViz (Eisenstein et al., 2012), and the topic visualization of (Chaney and Blei, 2012) support topic-based corpus browsing and understanding. Visualizations of this type typically represent standard topic models as a *sea of word clouds*; the individual topics within the model are presented as an unordered set of word clouds — or something similar — of the top words for the topic[1] where word size is proportional to the probability of the word for the topic. A primary issue with word clouds is that they can hinder understanding (Harris, 2011) due to the fact that they lack information about the relationships between words. Additionally, topic model visualizations that display topics in a random layout can lead to a huge, inefficiently organized search space, which is not always helpful in providing a quick corpus overview or assisting the user to diagnose possible problems with the model.

The authors of Correlated Topic Models (CTM) (Lafferty and Blei, 2006) recognize the limitation of existing topic models to directly model the correlation between topics, and present an alternative algorithm, CTM, which models the correlation between topics discovered for a corpus by using a more flexible distribution for the topic proportions in the model. Topical n-gram models (TNG) (Wang et al., 2007) discover phrases in addition to topics. TNG is a probabilistic model which assigns words and n-grams based on surrounding context, instead of for all references in the corpus. These models independently account for the two limitations of statistical topic modeling discussed in this paper by modifying the underlying topic modeling algorithm. Our work aims to provide a low-cost method for incorporating this

---

[1]This varies, but typically is either the top 10 to 20 words or the number of words which hold a specific portion of the distribution weight.

information as well as visualizing it in an effective way. We compute summary statistics, term co-occurrence and topic covariance, which can be overlaid on top of any traditional topic model. As a number of application-specific LDA implementations exist, we propose a meta-technique which can be applied to any underlying algorithm.

We present a *relationship-enriched* visualization to help users explore topic models through word and topic correlations. We propose interactions to support user understanding, validation, and refinement of the models.

2 Group-in-a-box Layout for Visualizing a Relationship-Enriched Topic Model

Existing topic model visualizations do not easily support displaying the relationships between words in the topics and topics in the model. Instead, this requires a layout that supports intuitive visualization of nested network graphs. A group-in-a-box (GIB) layout (Rodrigues et al., 2011) is a network graph visualization that is ideal for our scenario as it is typically used for representing clusters with emphasis on the edges within and between clusters. The GIB layout visualizes sub-graphs within a graph using a Treemap (Shneiderman, 1998) space filling technique and layout algorithms for optimizing the layout of sub-graphs within the space, such that related sub-graphs are placed together spatially. Figure 1 shows a sample group-in-a-box visualization.

We use the GIB layout to visually separate topics of the model as groups. We implement each topic as a force-directed network graph (Fruchterman and Reingold, 1991) where the nodes of the graph are the top words of the topic. An edge exists between two words in the network graph if the value of the term co-occurrence for the word pair is above a certain threshold,[2] and the edge is weighted by this value. Similarly, the edges between the topic clusters represent the topic covariance metric. Finally, the GIB layout optimizes the visualization such that related topic clusters are placed together spatially. The result is a topic visualization where related words are clustered within the topics and related topics are clustered within the overall layout.

---

[2]There are a variety of techniques for setting this threshold; currently, we aim to display fewer, stronger relationships to balance informativeness and complexity of the visualization
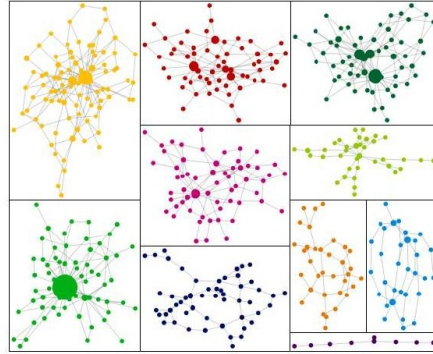
Figure 1: A sample GIB layout from (Rodrigues et al., 2011). The layout visualizes clusters distributed in a treemap structure where the partitions are based on the size of the clusters.

3 Relationship Metrics

We compute the term and topic relationship information required by the GIB layout as term co-occurrence and topic covariance, respectively. Term co-occurrence is a corpus-level statistic that can be computed independently from the LDA algorithm. The results of the LDA algorithm are required to compute the topic covariance.

3.1 Corpus-Level Term Co-Occurrence

Prior work has shown that Pointwise Mutual Information (PMI) is the most consistent scoring method for evaluating topic model coherence (Newman et al., 2010). PMI is a statistical technique for measuring the association between two observations. For our purposes, PMI is used to measure the correlation between each term pair within each topic on the document level[3]. The PMI is calculated for every possible term pair in the ingested data set using Equation 1. The visualization uses only the PMI for the term pairs for the top terms for each topic, which is a small subset of the calculated PMI values. Computing the PMI is trivial compared to the LDA calculation, and computing the values for all pairs allows the job to be run in parallel, as opposed to waiting for the results of the LDA job to determine the top term pairs.

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \quad (1)$$

The PMI measure represents the probability of observing x given y and vice-versa. PMI can be

---

[3]We use document here, but the PMI can be computed at various levels of granularity as required by the analyst intent.

positive or negative, where 0 represents independence, and PMI is at its maximum when x and y are perfectly associated.

3.2 Topic Covariance

To quantify the relationship between topics in the model, we calculate the topic covariance metric for each pair of topics. To do this, we use the theta vector from the LDA output. The theta vector describes which topics are used for which documents in the model, where theta(d,i) represents how much the ith topic is expressed in document d. The equations for calculation the topic covariance are shown below.

$$\gamma_{di} = \frac{\theta_{di}}{\sum_j(\theta_{dj})} \tag{2}$$

$$\gamma_i = \frac{1}{D}\sum_d(\gamma_{di}) \tag{3}$$

$$\sigma(i,j) = \frac{1}{D}\sum_d(\gamma_{di} - \gamma_i)(\gamma_{dj} - \gamma_j)) \tag{4}$$

4 Visualization

The visualization represents the individual topics as network graphs where nodes represent terms and edges represent frequent term co-occurrence, and the layout of the topics represents topic covariance. The most *connected* topic is placed in the center of the layout, and the least connected topics are placed at the corners. Figure 2 shows the visualization for a topic model generated for a 1,000 document NSF dataset. As demonstrated in Figure 3, a user can hover over a topic to see the related topics[4]. In this example, the user has hovered over the {visualization, visual, interactive} topic, which is related to {user, interfaces}, {human, computer, interaction}, {design, tools}, and {digital, data, web} among others. Unlike other topical similarity measures, such as cosine similarity or a count of shared words, the topic covariance represents topics which are typically discussed together in the same documents, helping the user to discover semantically similar topics.

On the topic level, the size of the node in the topic network graph represents the probability of the word given the topic. By mapping word probability to the area of the nodes instead of the height

---

[4]we consider topics related if the topic co-occurrence is above a certain pre-defined threshold.

Figure 2: The visualization utilizes a group-in-a-box-inspired layout to represent the topic model as a nested network graph.

of words, the resulting visual encoding is not affected by the length of the words, a well-known issue with word cloud presentations that can visually bias longer terms. Furthermore, circles can overlap without affecting a user's ability to visually separate them, and lead to more compact and less cluttered visual layout. Hovering over a word node highlights the same word in other topics as shown in Figure 4.

This visualization is an alternative interface for Interactive Topic Modeling (ITM) (Hu et al., 2013). ITM presents users with topics that can be modified as appropriate. Our preliminary results show that topics containing highly-weighted sub-clusters may be candidates for splitting, whereas positively correlated topics are likely to be *good* topics, which do not need to be modified. In future work, we intend to perform an evaluation to show that this visualization enhances quality and efficiency of the ITM process.

To support user interactions required by the ITM algorithm, the visualization has an edit mode, which is shown in Figure 5. Ongoing work includes developing appropriate visual operations to support the following model-editing operations:

1. Adding words to a topic
2. Removing words from a topic
3. Requiring two words to be linked within a topic (must link)
4. Requiring two words to be forced into separate topics (cannot link)

5 Conclusion and Future Work

The visualization presented here provides a novel way to explore topic models with incorporated

Figure 3: The user has hovered over the most-central topic in the layout, which is the most connected topic. The hovered topic is outlined, and the topic name is highlighted in turquoise. The topic names of the related topics are also highlighted.
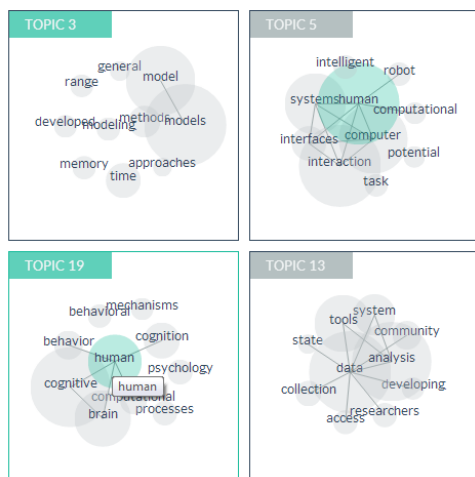

Figure 4: The visualization where the user has hovered over a word of interest. The same word is highlighted turquoise in other topics.
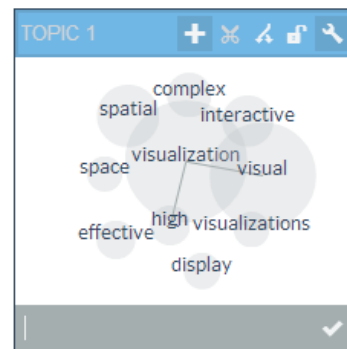

Figure 5: The edit mode for the visualization. From this mode, the user can add words, remove words, or rename the topic.

term and topic correlation information. This is a work in progress with the end goal to combine the visual representation with interactive topic modeling to allow users to explore (a) why a model may not align with their intuition and (b) modify the model as needed. We plan to deploy the tool on real-world domain users to iteratively refine the visualization and evaluate it in ecologically valid settings.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Machine Learning Journal*, 3:993–1022.

Allison June-Barlow Chaney and David M Blei. 2012. Visualizing topic models. In *ICWSM*.

Jacob Eisenstein, Duen Horng Chau, Aniket Kittur, and Eric Xing. 2012. Topicviz: interactive topic exploration in document collections. In *CHI'12 Extended Abstracts*, pages 2177–2182. ACM.

Thomas MJ Fruchterman and Edward M Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164.

Matthew J Gardner, Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi. 2010. The topic browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization*.

Jacon Harris. 2011. Word clouds considered harmful. http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2013. Interactive topic modeling. *Machine Learning*, pages 1–47.

JD Lafferty and MD Blei. 2006. Correlated topic models. In *NIPS, Proceedings of the 2005 conference*, pages 147–155. Citeseer.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *HLT*, pages 100–108. ACL.

Eduarda Mendes Rodrigues, Natasa Milic-Frayling, Marc Smith, Ben Shneiderman, and Derek Hansen. 2011. Group-in-a-box layout for multifaceted analysis of communities. In *ICSM*, pages 354–361. IEEE.

Ben Shneiderman. 1998. Treemaps for space-constrained visualization of hierarchies.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM*, pages 697–702. IEEE.