Termite: Visualization Techniques for Assessing Textual Topic Models

Jason Chuang, Christopher D. Manning, Jeffrey Heer
Stanford University Computer Science Department
{jcchuang, manning, jheer}@cs.stanford.edu

ABSTRACT

Topic models aid analysis of text corpora by identifying latent topics based on co-occurring words. Real-world deployments of topic models, however, often require intensive expert verification and model refinement. In this paper we present Termite, a visual analysis tool for assessing topic model quality. Termite uses a tabular layout to promote comparison of terms both within and across latent topics. We contribute a novel saliency measure for selecting relevant terms and a seriation algorithm that both reveals clustering structure and promotes the legibility of related terms. In a series of examples, we demonstrate how Termite allows analysts to identify coherent and significant themes.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.5.2 [**Information Interfaces**]: User Interfaces

General Terms

Algorithms, Design, Human Factors

Keywords

Topic Models, Text Visualization, Seriation

1. INTRODUCTION

Recent growth in text data affords an opportunity to study and analyze language at an unprecedented scale. The size of text corpora, however, often exceeds the limit of what a person can read and process. While statistical topic models have the potential to aid large-scale exploration, a review of the literature reveals a scarcity of real world analyses involving topic models. When the models are deployed, they involve time-consuming verification and model refinement.

We present Termite, a visualization system for the term-topic distributions produced by topic models. Our system contributes two novel techniques to aid topic model assessment. First, we describe a **saliency measure** for ranking and filtering terms. By surfacing more discriminative terms, our measure enables faster assessment and comparison of topics. Second, we introduce a **seriation method** for sorting terms to reveal clustering patterns. Our technique has two desirable properties: preservation of term reading order and early termination when sorting subsets of words. We demonstrate how these techniques enable rapid classification of coherent or junk topics and reveal topical overlap.

AVI '12, May 21-25, 2012, Capri Island, Italy

2. RELATED WORK

Latent Dirichlet allocation (LDA) [3] is a popular approach for uncovering *latent topics*: multinomial probability distributions over terms, generated by soft clustering of words based on document co-occurrence. While LDA produces some sensible topics, a prominent issue is the presence of "junk topics" [1] comprised of incoherent or insignificant term groupings. Model outputs often need to be verified by domain experts and modified [5] to ensure they correspond to meaningful concepts in the domain of analysis.

Hall et al. [12] applied LDA to study research trends in computational linguistics across 14,000 publications. The authors recruited experts to validate the quality of the latent topics. These experts retained only 36 out of 100 topics, and manually inserted 10 additional topics not produced by the model. Talley et al. [24] examined 110,000 NIH grants and applied LDA to uncover 700 latent topics. The modeling process included a significant amount of revision: modifying the vocabulary to include acronyms and multi-word phrases, removing nonsensical topics, conducting parameter search, and comparing the resulting models.

Current evaluations of topical quality rely heavily on experts examining lists of the most probable words in a topic [4, 19, 20]. For example, in biological texts one might find a topic with terms "*dna, replication, rna, repair, complex, interaction, ...*" Prior work in visualization suggests some alternative forms of presentation. Matrix views can surface relationships among a large number of items [2, 14] or between two data dimensions [9] if an appropriate ordering (or *seriation*) is applied [10, 26]. Interaction might then allow users to explore alternative orderings [22]. An appropriate model of words (e.g., statistically significant instead of frequent terms, phrases instead of words) can further aid comparison [7, 27]. Incorporating word relatedness into a visualization can surface high-level patterns in the text [6, 13]. In contrast to existing tools for summarizing LDA model output [11], Termite aims to support the domain-specific task of building and refining topic models.

3. THE TERMITE SYSTEM DESIGN

When using topic models to analyze a text collection, it is critical that the discovered latent topics be relevant to the domain task. Prior work suggests that the quality of a topic is often determined by the coherence of its constituent words [1] and its relative importance to the analysis task [25] in comparison to other topics. Effective means for assessing

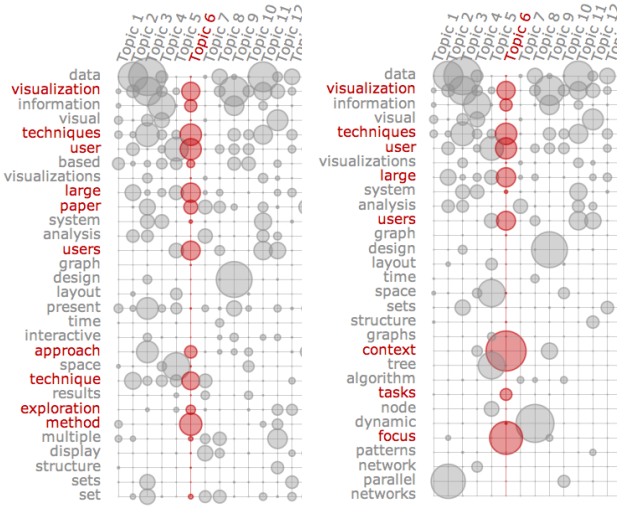

Figure 1: Top 30 frequent (left) vs. salient (right) terms. Our saliency measure ranks *tree*, *context*, *tasks*, *focus*, *networks* above the more frequent but less informative words *based*, *paper*, *approach*, *technique*, *method*. Distinctive terms enable speedier identification: Topic 6 concerns focus+context techniques; this topical composition is ambiguous when examining the frequent terms.

topical quality are thus an important step toward making topic models more useful for real-world analyses.

Our goal with Termite is to support *effective evaluation of term distributions associated with LDA topics*. The tool is designed to help assess the quality of individual topics and all topics as a whole. The primary visualization used in Termite is a matrix view; rows correspond to terms and columns to topics. In the following examples we use LDA models [21] with 25 to 50 topics, trained on abstracts from 372 IEEE InfoVis conference papers from 1995 to 2010 [23].

The **term-topic matrix** (Figures 1–3) shows term distributions for all latent topics. Unlike lists of per-topic words (the current standard practice), matrices support comparison across both topics and terms. We use circular area to encode term probabilities. Texts typically exhibit long tails of low probability words. Area has a higher dynamic range than length encodings (quadratic vs. linear scaling) and curvature enables perception of area even when circles overlap. We also experimented with parallel tag clouds [7] where text is displayed directly in the matrix; the result was not sufficiently compact for even a modest number of terms.

Users can **drill down** to examine a specific topic by clicking on a circle or topic label in the matrix. The visualization then reveals two additional views. The word frequency view (Figure 3, middle) shows the topic's word usage relative to the full corpus. The document view (Figure 3, right) shows the representative documents belonging to the topic.

3.1 Displaying Informative Terms

Showing all words in the term-topic matrix is neither desirable nor feasible due to large vocabularies with thousands of words. Termite can **filter** the display to show the most *probable* or *salient* terms. Users can choose between 10 and 250 terms. On most monitors displaying over 250 words requires a significant amount of scrolling and reduces the effectiveness of the visualization.

Table 1: Word similarity based on G^2 statistics

G^2 estimates the likelihood of an event v taking place when another event u is also observed. The likelihood can be computed [8] using the following 2×2 contingency table:

events	u	$\neg u$
v	$a = P(u\|v)$	$b = P(\neg u\|v)$
$\neg v$	$c = P(u\|\neg v)$	$d = P(\neg u\|\neg v)$

The G^2 statistic is then defined as:

$$G^2 = a \log \frac{a(c+d)}{c(a+b)} + b \log \frac{b(c+d)}{d(a+b)}$$

For word co-occurrences, G^2 represents the likelihood of a word v appearing in a document/sentence when another word u also appears in the same document/sentence. For bigrams, G^2 examines all adjacent pairs of words, and estimates the likelihood of v being the second word when u is the first word.

We define **term saliency** as follows. For a given word w, we compute its conditional probability $P(T|w)$: the likelihood that observed word w was generated by latent topic T. We also compute the marginal probability $P(T)$: the likelihood that any randomly-selected word w' was generated by topic T. We define the *distinctiveness* of word w as the Kullback-Leibler divergence [15] between $P(T|w)$ and $P(T)$:

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

This formulation describes (in an information-theoretic sense) how informative the specific term w is for determining the generating topic, versus a randomly-selected term w'. For example, if a word w occurs in all topics, observing the word tells us little about the document's topical mixture; thus the word would receive a low distinctiveness score.

The *saliency* of a term is defined by the product:

$$saliency(w) = P(w) \times distinctiveness(w)$$

As shown in Figure 1, filtering terms by saliency can aid rapid classification and disambiguation of topics. Given the same number of words, the list of most probable terms contains more generic words (e.g., *based*, *paper*, *approach*) than the list of distinctive terms (e.g., *tree*, *context*, *tasks*). Our saliency measure speeds identification of topical composition (e.g., Topic 6 on focus+context techniques). By producing a more sparse term-topic matrix, our measure can enable faster differentiation among the topics and identification of potential "junk topics" lacking salient terms.

3.2 Ordering the Term-Topic Matrix

Termite provides two options for **topic ordering**: by *index* (the arbitrary topic index produced by LDA) and by *topic size* (the number of observed terms assigned to a topic). Prior work suggests that small (rare) topics tend to contain more nonsensical and incoherent terms [19]. Topic ordering by size can help surface such patterns.

Termite also provides three options for **term ordering**: *alphabetically*, by *frequency*, or using *seriation*. Seriation methods permute the presentation order to reveal clustering structure, and are commonly used to improve visualizations of matrices [16] and cluster heatmaps [10].

Termite uses a novel **seriation method for text data**. First, we define an *asymmetric similarity measure* to account for co-occurrence and collocation likelihood between all pairs of words. Collocation defines the probability that

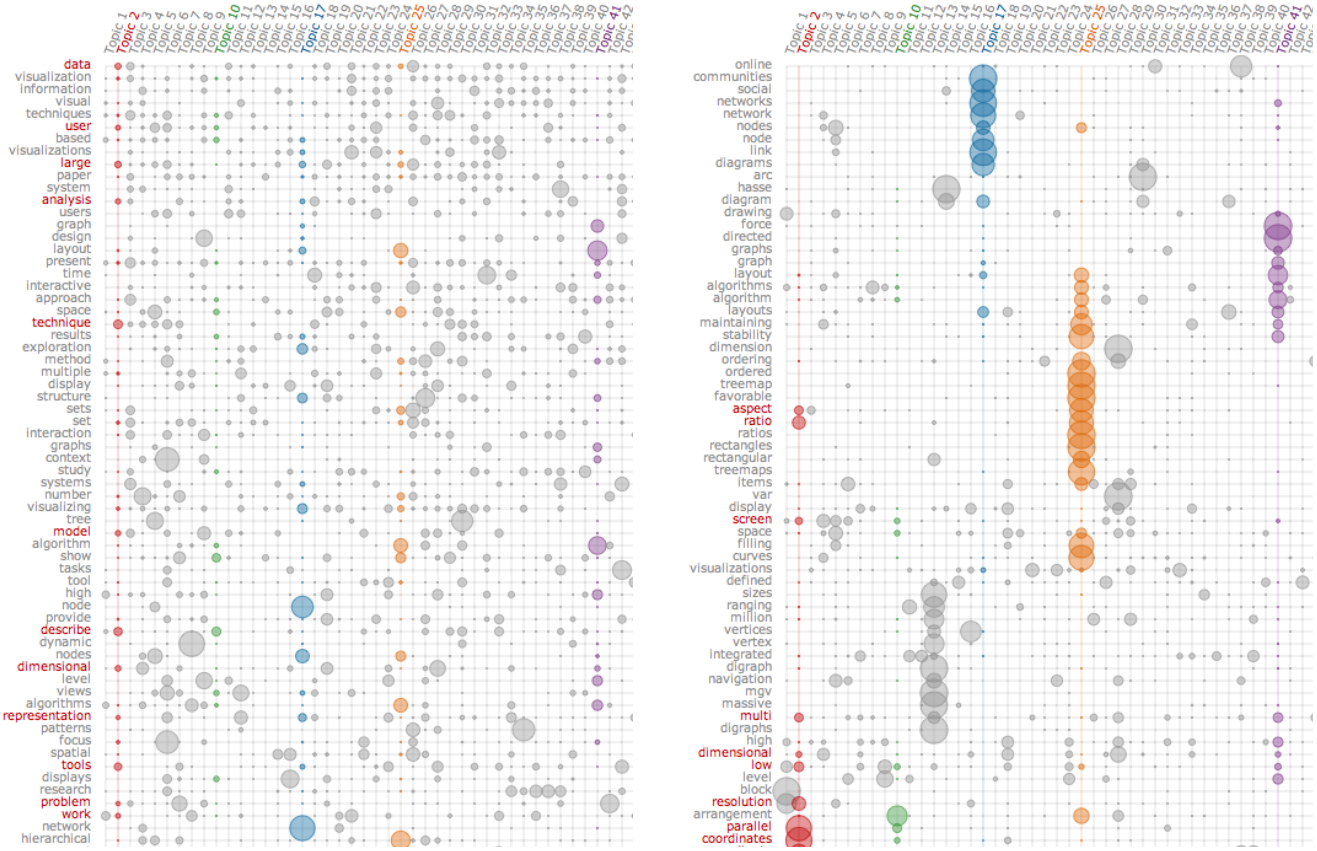

Figure 2: Seriation. Terms ordered by frequency (left) vs. our seriation technique (right). Seriation reveals clusters of terms and aids identification of coherent concepts such as Topic 2 (parallel coordinates), Topic 17 (network visualization), Topic 25 (treemaps), and Topic 41 (graph layout). Our term similarity measure embeds word ordering and favors reading order (*online communities*, *social networks*, *aspect ratio*, *etc*).

a phrase (sequence of words) occurs more often in a corpus than would be expected by chance, and is an asymmetric measure. For example, "social networks" is a likely phrase; "networks social" is not. Incorporating collocation favors adjacent words that form meaningful phrases, in the correct reading order. We compute the likelihoods using G^2 statistics [8] as shown in Table 1.

We then place the terms according to their similarity scores by applying the Bond Energy Algorithm [18]. We terminate BEA whenever a sorted sub-list with the desired number of terms is generated. Assessing topical composition typically requires examining only a subset of the common or mid-frequency words [17], and does not require seriating the full vocabulary. We use BEA because it accepts asymmetric similarity measures as input and is a greedy algorithm; early termination does not affect the quality of its results.

As shown in Figure 2, our seriation algorithm reveals topical clusters of terms. We are able to rapidly identify coherent concepts such as Topic 2 on parallel coordinates. Term grouping reveals shared properties between topics, e.g., *maintaining stability* in both treemaps and force-directed graph layout. Our technique preserves reading order down the list of terms; examples include *online communities*, *social networks* and *aspect ratio*. Seriating terms in reading order facilitates scanning and a sense of term use in context.

Qualitatively, we observe that seriating terms using a combined similarity measure based on both document and sentence level co-occurrence is preferable to either statistics alone. Bigram likelihood produces a significantly sparser matrix than does document co-occurrence alone. As a result, adding bigram likelihood doesn't significantly change the global seriation order. Instead, it affects local orderings and places words such as *parallel coordinates*, *user interface*, *social networks* and *small multiples* in the correct reading order. We experimented with trigram statistics, but find that it degrades the overall seriation quality. Longer phases such as *node link diagram* are already produced by bigram statistics. Adding trigrams yields marginal gains and produces phrases such as *graph layout algorithm*, *large data set*, and *social network analysis*. However, adding trigram likelihood leads to false positives: because the stop word *of* is omitted, the recurring trigram *level of detail* adds undesirable weight to the word sequence *level detail*.

4. CONCLUSION

Based on usage by members of our research group, we observed that users are able to meaningfully comprehend topical composition with Termite. Example quotes include: *"The current [dataset] seems to overfit in places... much more so than the 30 topic example I used in [a previous iteration]"* and *"We may have single-doc topics!"*. We also received initial feedback requesting the ability to label and organize topics and examine document-topic probabilities.

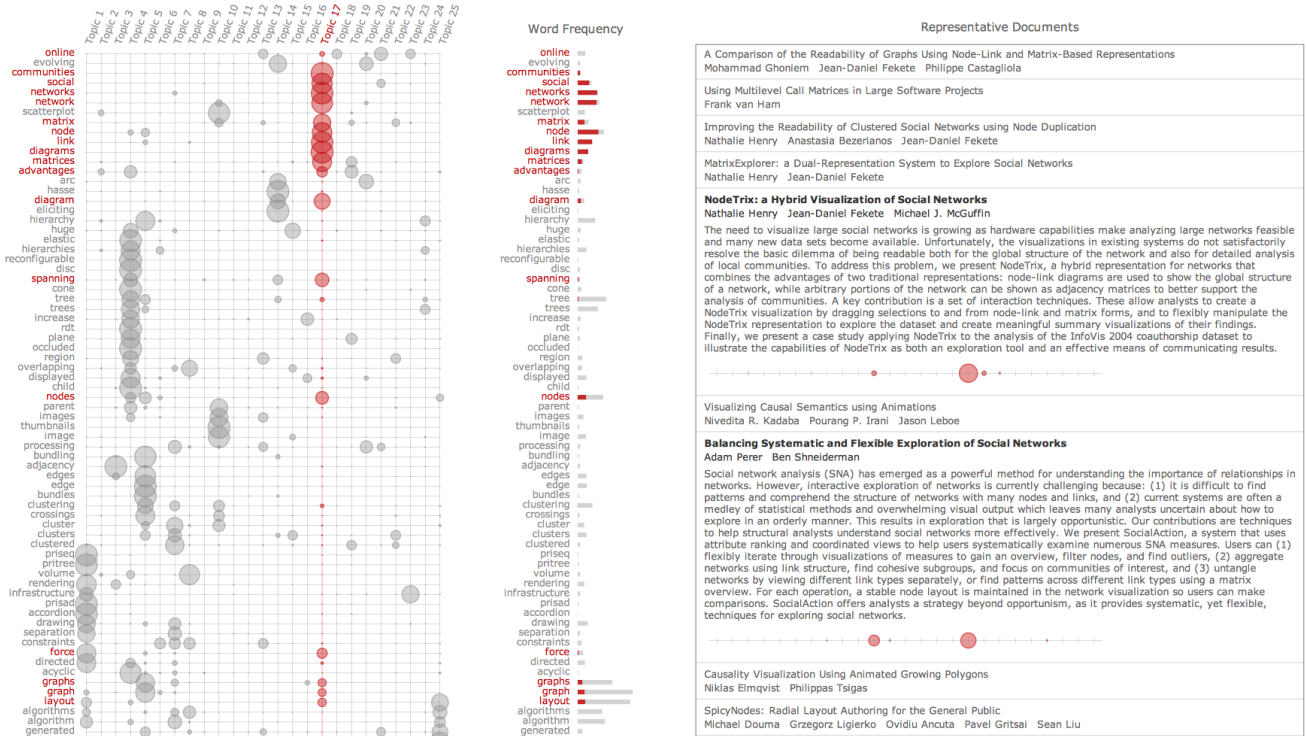

Figure 3: The Termite system. When a topic is selected in the term-topic matrix (left), the systems visualizes the word frequency distribution relative to the full corpus (middle) and shows the most representative documents (right).

Going forward, Termite is a first step towards a visual analysis system for human-centered iterative topic modeling. In this paper, we focused on understanding terms and term-topic distributions. Future work involves expanding Termite to visualize the topical composition of documents and adding interactions to support user inputs (e.g., adjusting model parameters, deleting junk topics, merging related topics). We believe supporting interactive model refinement can significantly improve the utility and reduce the cost of applying topic models to make sense of large text corpora.

5. REFERENCES

[1] L. Alsumait, D. Barbará, J. Gentle, and C. Domeniconi. Topic significance ranking of LDA generative models. In *ECML*, 2009.

[2] J. Bertin. Semiology of Graphics: Diagrams, Networks, Maps. University of Wisconsin Press, 1983.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J Machine Learning Research*, 3:993–1022, 2003.

[4] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, pages 288–296, 2009.

[5] J. Chuang, C. D. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *CHI*, 2012.

[6] C. Collins, S. Carpendale, and G. Penn. DocuBurst: Visualizing document content using language structure. *Computer Graphics Forum*, 28(3):1039–1046, 2009.

[7] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *VAST*, pages 91–98, 2009.

[8] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[9] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.

[10] M. Friendly. The history of the cluster heat map. *The American Statistician*, 2009.

[11] M. J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppi. The Topic Browser: An interactive tool for browsing topic models. In *NIPS*, 2010.

[12] D. Hall, D. Jurafsky, and C. D. Manning. Studying the history of ideas using topic models. In *EMNLP*, pages 363–371, 2008.

[13] M. A. Hearst. TileBars: visualization of term distribution information in full text information access. In *CHI*, 1995.

[14] N. Henry and J.-D. Fekete. MatLink: enhanced matrix visualization for analyzing social networks. In *Interact*, 2007.

[15] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[16] I. Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining*, 3:70–91, 2010.

[17] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[18] W. T. McCormick, P. J. Schweitzer, and T. W. White. Problem decomposition and data reorganization by a clustering technique. *Operations Research*, 20(5):993–1009, 1972.

[19] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *EMNLP*, pages 262–272, 2011.

[20] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin. Evaluating topic models for digital libraries. In *JCDL*, 2010.

[21] D. Ramage. Stanford topic modeling toolbox. http://nlp.stanford.edu/software/tmt/tmt-0.4.

[22] R. Rao and S. K. Card. The Table Lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In *CHI*, 1994.

[23] J. Stasko, C. Görg, Z. Liu, and K. Singhal. Jigsaw: Supporting investigative analysis through interactive visualization. In *VAST*, pages 131–138, 2007.

[24] E. M. Talley, D. Newman, D. Mimno, B. W. Herr, H. M. Wallach, G. A. P. C. Burns, A. G. M. Leenders, and A. McCallum. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, 8(6):443–444, 2011.

[25] Z. Wen and C. Yung Lin. Towards finding valuable topics. In *ICDM*, pages 720–731, 2010.

[26] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[27] K. Yatani, M. Novati, A. Trusty, , and K. N. Truong. Review Spotlight: A user interface for summarizing user-generated reviews using adjective-noun word pairs. In *CHI*, 2011.