SemVis: Semantic Visualization for Interactive Topical Analysis

Tuan M. V. Le
Singapore Management University
80 Stamford Road
Singapore 178902
vmtle.2012@phdis.smu.edu.sg

Hady W. Lauw
Singapore Management University
80 Stamford Road
Singapore 178902
hadywlauw@smu.edu.sg

ABSTRACT

Exploratory analysis of a text corpus is an important task that can be aided by informative visualization. One spatially-oriented form of document visualization is a scatterplot, whereby every document is associated with a coordinate, and relationships among documents can be perceived through their spatial distances. Semantic visualization further infuses the visualization space with latent semantics, by incorporating a topic model that has a representation in the visualization space, allowing users to also perceive relationships between documents and topics spatially. We illustrate how a semantic visualization system called SemVis could be used to navigate a text corpus interactively and topically via browsing and searching.

KEYWORDS

Semantic visualization; topic model; interactive topical analysis

1 INTRODUCTION

There are tasks that involve exploration of a text corpus for understanding of the corpus and extracting specific information. E.g., a scientist conducts literature review, a financial analyst digests economic reports, a patent officer examines prior art, a legal researcher looks for precedence. These scenarios involve various information needs, e.g., what the corpus is about in general, what the predominant topics are, which documents are relevant to a particular search intent, which other documents are related to the current document.

The original representation of a document is often a bag of words. It is high-dimensional, with dimensionality equal to the size of the vocabulary. One way to visualize document relationships is to reduce their high-dimensional representation into a low-dimensional one that preserves their similarities [5, 10]. Each document is associated with a coordinate in a 2D or 3D scatterplot. Similarities among documents can be perceived spatially via their close distances.

Such a visualization, on its own, is not designed for revealing the main "themes" of a corpus. A topic model [1, 7] associates each document with a probability distribution over topics, where the semantics can be interpreted by each topic's word distribution. It is common to model tens to hundreds of topics in a corpus, thus the topical dimensionality is still too high to be visualized directly.

CIKM'17, November 6–10, 2017, Singapore.

ISBN 978-1-4503-4918-5/17/11…$15.00
DOI: https://doi.org/10.1145/3132847.3133181

Our objective is to infuse the visualization with latent semantics. Recent developments in *semantic visualization* [9, 11, 12] jointly model topics and visualization coordinates. In this paradigm, documents and topics are respectively associated with latent visualization coordinates. A document's topic distribution is a function of the relative distance between the document's coordinate and each topic's coordinate. As a result, we can visualize the relationship not only between a pair of documents, but also between a document and a topic. The visualization space is also a continuous semantic space, as every coordinate (even an empty spot) codes for a distribution over topics, and by extension also a distribution over words, lending semantic interpretation to any point in the visualization.

Contributions and Organization. In this demo, we showcase SemVis, a semantic visualization system for interactive topical analysis. This is a demonstrable system that is built on, and is generically compatible with previous algorithmic works on semantic visualization [9, 11, 12] reviewed in Section 2. We illustrate the interactive topical analysis features and the capabilities of SemVis in browsing and searching scenarios in Section 3, and describe a pilot user study in Section 4. We briefly outline the implementation in Section 5, and describe the various scenarios supported by the demo in Section 6.

Related Work. To our best awareness, SemVis is the first demonstrable semantic visualization system. There exist other visualizations for text analysis, some of which involve topic modeling, but they are not oriented towards spatially-based semantic visualization. [2] presents a list-based interface, showing the text content of a single document and listing its topics. [3] sports a matrix-based interface, showing the important words for each topic. [14] tracks the relative strength of topics over time. [6] is based on network visualization. [13] is interested in how topics are related, rather than on how documents are related to each other and to topics.

2 SEMANTIC VISUALIZATION

Recent works on semantic visualization [9, 11, 12] jointly model topics and visualization coordinates. These are compatible with SemVis' visualization. We use SEMAFORE[12], which is one of the state-of-the-art models. For a brief review, SEMAFORE assumes that each document d_n and topic z have coordinates x_n and ϕ_z respectively. The generative model for a corpus is as follows.

(1) For each topic $z = 1, \ldots, k$:
  (a) Draw z's distribution of words: $\theta_z \sim \text{Dirichlet}(\alpha)$
  (b) Draw z's coordinate: $\phi_z \sim \text{Normal}(0, \beta^{-1}I)$
(2) For each document d_n where $n = 1, ..., N$:
  (a) Draw d_n's coordinate: $x_n \sim \text{Normal}(0, \gamma^{-1}I)$
  (b) For each word $w_{nm} \in d_n$:
    (i) Draw a topic: $z \sim \text{Multi}(\{\text{P}(z|x_n, \Phi)\}_{z=1}^{k})$
    (ii) Draw a word: $w_{nm} \sim \text{Multi}(\theta_z)$

α is a Dirichlet prior, I is an identity matrix, β and γ control the variance of the Normal distributions. $\mathrm{P}(z|x_n, \Phi)$ defines how the coordinate of each document x_n transforms into its topic distribution, according to Equation 1. The closer is x_n to the topic coordinate ϕ_z, the higher is the probability. Φ is the collection of topic coordinates.

$$\mathrm{P}(z|x_n, \Phi) = \frac{\exp(-\frac{1}{2}||x_n - \phi_z||^2)}{\sum_{z'=1}^{k} \exp(-\frac{1}{2}||x_n - \phi_{z'}||^2)} \quad (1)$$

The log likelihood function is shown in Equation 2.

$$\mathcal{L} = \sum_{n=1}^{N} \sum_{m=1}^{M_n} \log \sum_{z=1}^{k} \mathrm{P}(z|x_n, \Phi)\mathrm{P}(w_{nm}|\theta_z) \quad (2)$$

To preserve the locality structure of documents, the method employs neighborhood regularization as in Equation 3. λ is the regularization parameter. y_{ij} encodes the neighborhood graph, with $y_{ij} = 1$ signifying that d_i and d_j are neighbors, and $y_{ij} = 0$ otherwise. The regularized $\mathbf{L}$ reflects the idea that similar (neighboring) documents should have closer coordinates, while different (non-neighboring) documents should have further coordinates.

$$\mathbf{L} = \mathcal{L} - \frac{\lambda}{2}\left[\sum_{\substack{i,j=1 \\ i \neq j}} y_{ij}||x_i - x_j||^2 + \sum_{\substack{i,j=1 \\ i \neq j}} \frac{1 - y_{ij}}{||x_i - x_j||^2 + 1}\right] \quad (3)$$

The parameters are learned based on maximum a posteriori estimation through EM [4]. The outputs are the coordinate x_n, as well as distribution over k topics $\{\mathrm{P}(z|x_n, \Phi)\}_{z=1}^{k}$, for every document.

3 INTERACTIVE TOPICAL ANALYSIS

We describe the features of the visualization system SemVis, assuming that the coordinates and the topic distributions have been learned from the corpus as in the previous section. For the running example, we use a corpus based on $20News$[1] and learn 30 topics.

Browsing. Figure 1 shows the main screen of SemVis. Item (1) is the black canvas space for displaying the visualization. In this canvas, we display a 2D scatterplot of documents and topics. Each document is a circle. Each topic is a square, and is associated with a color. Item (2) is a legend of topics, listing the top words with the highest probabilities for each topic to aid topic intrepretation. While a document's coordinate codes for a probability distribution over all topics, for ease of identification, a document is colored the same as the topic with the largest probability in that document.

The layout as well as the coloring of documents and topics in the canvas reveal an overview of the corpus, in terms of the various topics that are relevant to the corpus, as well as the relationship among documents. We can perceive when documents are similar, both through their close distances as well as similar colors. Each cluster also tends to be "anchored" by a topic. Intuitively, documents in between two topics have significant probabilities for both.

For a detailed view, users can click on a circle (document) to see its content displayed on item (3), and its topic distribution on item (4). A list of interactive functions are provided on item (5), including zooming in and out of the canvas to focus on a specific region.

Every point x in the visualization space is associated with a topic distribution $P(z|x, \Phi)$ (see Equation 1). Taking into account each topic z's distribution over words $P(w|\theta_z)$, the point's word distribution can be obtained via $P(w|x) = \sum_z P(w|\theta_z)P(z|x, \Phi)$. At any point in this visualization space, the user can right-click to see the distribution of words corresponding to that point in space. For an example, item (6) in Figure 1 shows the list of top words associated with the coordinate on the top left corner of the list.

Searching. Other than browsing for a general understanding, users may need to search for a relevant set of documents. A traditional search engine returns a ranked list. While this is a familiar interface to search users, there are some aspects for which a visualization could be beneficial. For one, a query may be ambiguous, with a few different senses, e.g., "apple" the company vs. "apple" the fruit. A ranked list frequently interleaves results of different senses. For another, results within the ranked list may have a natural clustering structure, e.g., news about the same event. The ranking by relevance alone may not capture this, requiring additional processing.

Figure 2 shows our search interface. Currently, we support two query types. The first type is *textual query*. User can type in a query, and the most relevant results[2] are returned and displayed on a 2D visualization space. Here, we indicate the degree of relevance by the size of the circle, i.e., a more relevant document is drawn as a larger circle. The left panel of Figure 2 shows an example query *"fast drive"*. We can see clearly three clusters of results: a red cluster on the bottom right, a green cluster at the centre, and a blue cluster on the top left. This reveals that the query is indeed ambiguous, and it can be associated with several topics or senses. The red topic is about *card, problem, scsus, drive*, suggesting that the query is probably interpreted as about a fast driver software for some computer component. The green topic is about *system, disk, mac, software*, pointing to a fast hard disk drive. The blue topic is about *car, article, write, bike*, implying a fast driving car or bike.

The user may wish to refine the query to find more documents of a particular sense or topic. This is where the second query type, *spatial query*, may be useful. The user can specify any coordinate, and we return the "most relevant" or the closest documents within a radius. Continuing the example, if the user decides to focus on any one cluster, she can execute a spatial query by double-clicking a specific coordinate. On the right-hand side of Figure 2, we show three small panels, illustrating the hypothetical scenarios in which the user is interested in one of the three localities. Each panel corresponds to a spatial query, centered at the coordinate marked with an 'x'. This is akin to a visual interface for query reformulation.

4 PILOT USER STUDY

We conduct a pilot user study on $20News$ to confirm that semantic visualization with infused semantics is effective for users to perceive relationships between documents and topics. We compare three types of visualization. The first type is semantic visualization generated by Semafore where documents are colored based on their topics, and representative words for each topic is displayed as shown in Figure 1. The second type (Raw Semafore) where we use only document coordinates and remove topic information (colors and representative words). The third type is traditional scatterplot generated from t-SNE [5], which does not have topic information.

We design the user study as follows. For each question, we use an ambiguous word to search for documents. A visualization of the most relevant 20 documents (cosine similarity on tf-idf vectors) is

[1] http://ana.cachopo.org/datasets-for-single-label-text-categorization

[2] We return up to 50 most relevant results, which is a configurable number.

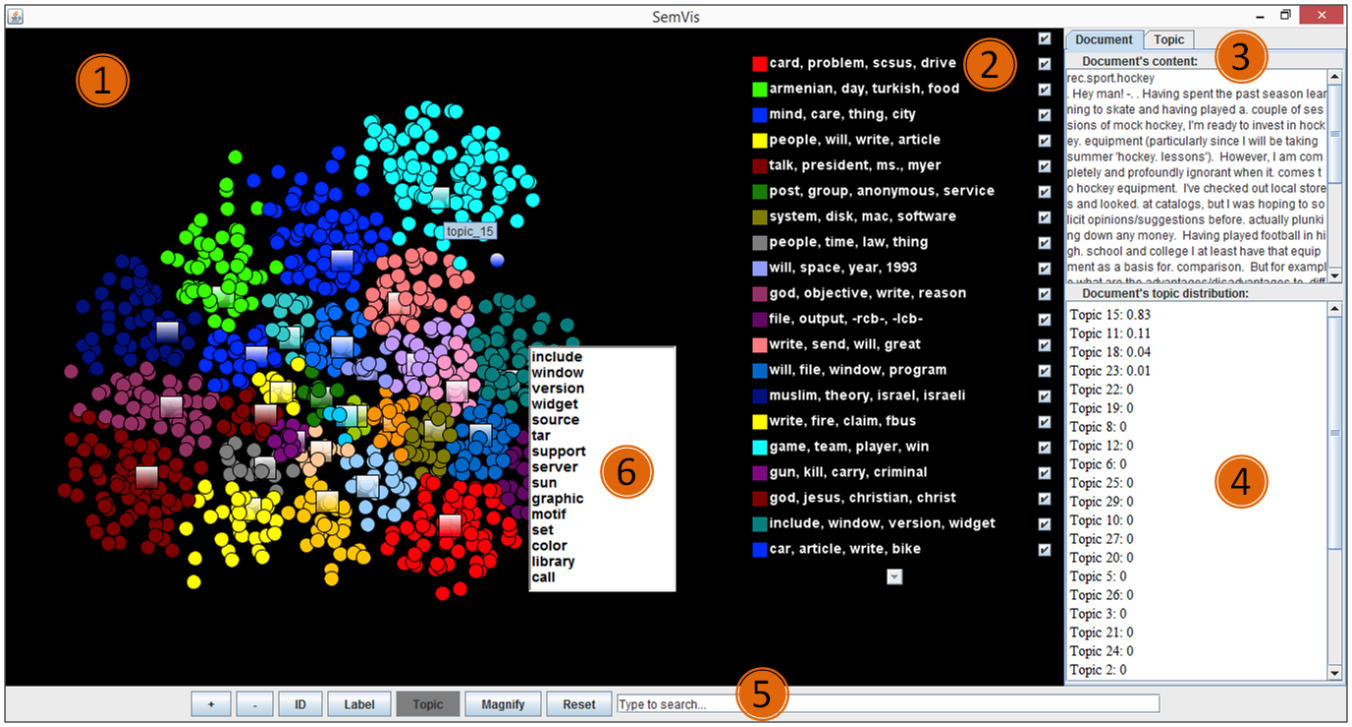

Figure 1: Browsing Interface

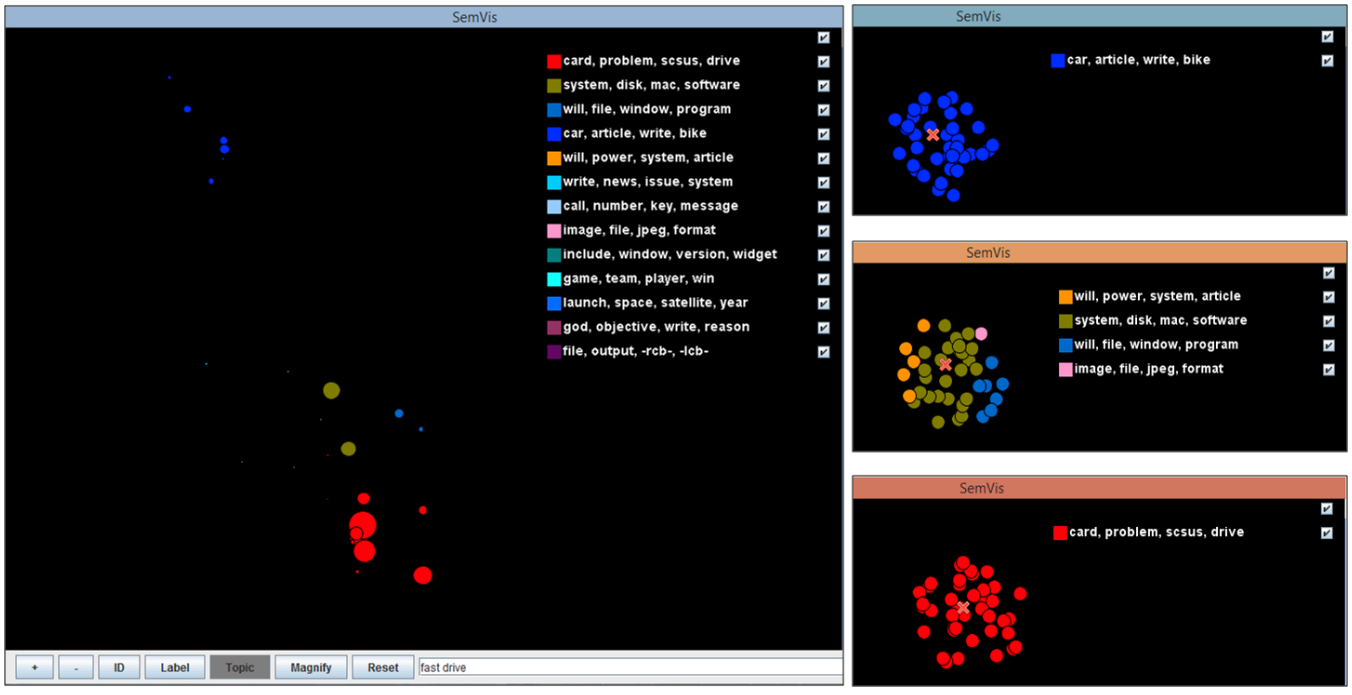

Figure 2: Search Interface: text query (left) and spatial queries (right). Topics of retrieved documents are shown in the legend.

	Precision (%)	Recall (%)	F1 (%)	Time (s)
Semafore	**43.7**	**61.0**	**50.9**	**77.3**
Raw Semafore	39.8	51.5	44.9	84.0
t-SNE	37.2	44.0	40.3	82.9

Table 1: Results of the user study (bold is better)

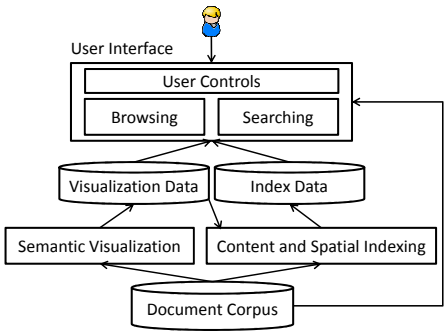

Figure 3: Framework of SemVis

presented to users. We ask users to find documents belonging to a specific category within a time limit of 90 seconds. There are 20 such ambiguous words[3], selected from the vocabulary, with the highest entropy computed based on categories of returned documents. For each question, the visualization is generated either by Semafore, Raw Semafore or t-SNE. Each user is randomly presented with one of the three versions. The users do not know how many methods there are. There are 6 users (who are not the authors) in the study.

Table 1 summarizes the results of the user study. When comparing to t-SNE, semantic visualization represented by Semafore helps users to achieve better precision (43.7% vs. 37.2%), recall (61% vs. 44%), and F1 (50.9% vs. 40.3%), while needing less time to complete the tasks (77.3s vs. 82.9s). This showcases that users can better perceive the document similarities and topics through semantic visualization. Comparing Semafore to Raw Semafore, we see that without infused topics (i.e., colors and topic representative words) users' performance drops (e.g., from 43.7% to 39.8% for precision). This showcases the importance of topics that are indirectly displayed by colors and topic representative words. Finally, although the time needed using Raw Semafore is longer (about 1 second), users can attain higher precision, recall, and F1 comparing to t-SNE. This performance gain can be explained by the fact that document distances in Semafore also reflect the semantic similarities of documents, which help users to determine similar documents.

5 IMPLEMENTATION

We briefly discuss the implementation. Figure 3 shows the framework of SemVis. It has three main modules. The first module, *Semantic Visualization*, helps to build a topic model and visualization of the corpus. We use Semafore [12][4], but the framework is compatible with other algorithms such as PLSV [9] or SSE [11], or even pipelines of a topic model, e.g., LDA [1], followed by embedding, e.g., PE [8]. The second module, *Content and Spatial Indexing*, provides functions for indexing the corpus. We use Apache Lucene 6.4.1[5] implemented in Java. We index two kinds of information: text content and visualization coordinates. The third module, *User Interface*, provides controls for performing browsing and searching easily, such as dragging, selecting, zooming and magnification, as well as a search box. We rely on Jung Library 2.0.1[6] written in Java.

6 DEMO

Data. For demonstration, we rely on several English corpora. One is $20News$[1], which is also used to provide the illustrations in this paper. We also rely on several text corpora obtained from *Cora*[7], which is a collection of abstracts of academic publications from various categories. From *Cora*, we carve out four smaller text corpora based on categories, namely *Data Structure* with 570 documents, *Hardware and Architecture* with 223 documents, *Machine Learning* with 1980 documents, and *Programming Language* with 1553 documents.

Scenario. We will allow the audience to freely interact with the system through the various browsing controls and interactive querying (for any search query suitable for the corpus at hand).

SemVis is a demonstrable system for interactive topical analysis via spatial visualization, supported by the rigor of the underlying semantic visualization algorithms in deriving topics and coordinates. Through the demo, we hope to spark a continuing conversation on the applicability of semantic visualization for text analysis tasks.

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR* 3 (2003), 993–1022.
[2] Allison June-Barlow Chaney and David M Blei. 2012. Visualizing Topic Models. In *ICWSM*.
[3] Jason Chuang, Christopher D Manning, and Jeffrey Heer. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 74–77.
[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *JRSS, Series B* 39, 1 (1977), 1–38.
[5] L. Van der Maaten and G. Hinton. 2008. Visualizing data using t-SNE. *JMLR* 9 (2008).
[6] Brynjar Gretarsson, John OâĂŹdonovan, Svetlin Bostandjiev, Tobias Höllerer, Arthur Asuncion, David Newman, and Padhraic Smyth. 2012. Topicnets: Visual analysis of large text corpora with topic modeling. *TIST* 3, 2 (2012), 23.
[7] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *SIGIR*. 50–57.
[8] Tomoharu Iwata, Kazumi Saito, Naonori Ueda, Sean Stromsten, Thomas L Griffiths, and Joshua B Tenenbaum. 2007. Parametric embedding for class visualization. *Neural Computation* 19, 9 (2007), 2536–2556.
[9] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. 2008. Probabilistic latent semantic visualization: topic model for visualizing documents. In *KDD*. 363–371.
[10] J. B. Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1 (1964).
[11] Tuan M V Le and Hady W Lauw. 2014. Semantic visualization for spherical representation. In *KDD*. 1007–1016.
[12] Tuan M V Le and Hady W Lauw. 2016. Semantic Visualization with Neighborhood Graph Regularization. *JAIR* 55 (2016), 1091–1133.
[13] Carson Sievert and Kenneth E Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. 63–70.
[14] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. Tiara: a visual exploratory text analytic system. In *KDD*. 153–162.

[3] The 20 queries are: air, ball, bank, base, beat, board, channel, chip, contact, crack, cross, drive, game, head, hook, match, patch, service, stick, strike.
[4] https://github.com/tuanlvm/SEMAFORE
[5] http://lucene.apache.org/core/6_4_1/
[6] http://jung.sourceforge.net/
[7] http://people.cs.umass.edu/~mccallum/data/cora-classify.tar.gz