9-2021

Semi-supervised semantic visualization for networked documents

Delvin Ce ZHANG
Singapore Management University, cezhang.2018@smu.edu.sg

Hady W. LAUW
Singapore Management University, hadywlauw@smu.edu.sg

Citation

ZHANG, Delvin Ce and LAUW, Hady W.. Semi-supervised semantic visualization for networked documents. (2021). *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17: Proceedings*. 12977, 762-778.
Available at: https://ink.library.smu.edu.sg/sis_research/6428

Semi-Supervised Semantic Visualization for Networked Documents

Delvin Ce Zhang$^{[0000-0001-5571-9766]}$ (✉) and Hady W. Lauw$^{[0000-0002-8245-8677]}$

School of Computing and Information Systems, Singapore Management University, Singapore
{cezhang.2018,hadywlauw}@smu.edu.sg

Abstract. Semantic interpretability and visual expressivity are important objectives in exploratory analysis of text. On the one hand, while some documents may have explicit categories, we could develop a better understanding of a corpus by studying its finer-grained structures, which may be latent. By inferring latent topics and discovering keywords associated with each topic, one obtains a semantic interpretation of the corpus. One the other hand, by visualizing documents, latent topics, and category labels on the same plot, one gains a bird's eye view of the relationships among documents, topics, and various categories. Semantic visualization is a class of methods that unify both topic modeling and visualization. In this paper, we propose a novel semantic visualization model for networked documents that incorporates partial labels. We introduce coordinate-based label distribution and label-dependent topic distribution to visualize documents, topics, and labels in a semi-supervised way. We further derive three variants for singly-labeled, multi-labeled, and hierarchically-labeled documents. The focus on semi-supervision that employs variants of labeling structures is particularly novel. Experiments verify the efficacy of our model against baselines.

Keywords: Semantic Visualization · Topic Modeling · Dimensionality Reduction · Generative Models.

1 Introduction

While text documents are mainly expressed in words, in many cases they are interconnected in a network, e.g., Web page hyperlinks or paper citations. When exploring such a corpus, we seek a comprehensive understanding in terms of both latent semantics and document proximity. On one hand, *topic modeling* excels at latent semantics. It represents each document by a topic distribution, and a topic is described by a group of keywords. Lacking visual interpretation, it requires cognitive efforts to summarize. *Visualization*, on the other hand, provides another view to understand the corpus by projecting high-dimensional documents to a low-dimensional space (2D or 3D), so similar documents could be found in spatial proximities. But it offers no lexical nor semantic interpretability. Given such tradeoffs, a promising direction is to pursue the 'joint' avenue of *semantic visualization*, which conducts topic modeling and visualization simultaneously, and visualizes documents and topics in the same scatterplot.

Existing semantic visualization models are mainly *unsupervised*. They do not take advantage of the fact that many corpora are partially labeled. Documents may be partitioned into categories, such as primary areas of academic publications. From this observation, we draw three critical insights. First, visualizing category labels in addition to documents and topics would better flesh out the corpus structure, as labels summarize a group of topics, and topics characterize documents. Second, by exploiting label structure, we could improve topic modeling, as documents within the same category would share topics or neighbors. Third, even partially available labels would be useful, if the modeling could induce probabilistic labels in a semi-supervised way. Note that labels are different from topics. The former capture a category of documents and are explicit and observed, the latter are completely latent. A document usually has more latent topics than observed labels. Documents of the same label may still vary in topics.

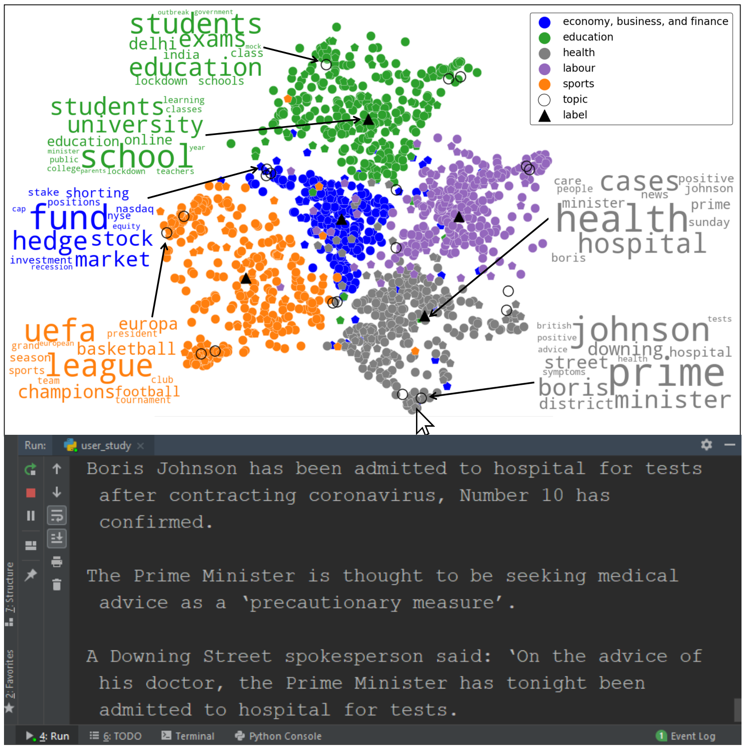
Of particular interest is the existence of several label structures. Single-label would be the most common, each document is assigned only one label. Alternatively, documents may also be tagged, giving rise to a multi-label structure, e.g., news articles with multiple tags. In some scenarios, the categorization may even be hierarchical, e.g., academic papers from the same area further fall into different sub-areas. We seek to design a semantic visualization model capable of accommodating different label structures.

The proposed model is called SemiVN, a **Semi**-supervised topic model for semantic **V**isualization of **N**etworked documents. The *first* key design is to introduce coordinate-based label distribution and label-dependent topic distribution to visualize documents, topics, and labels on the same scatterplot. One can infer how documents relate to topics and how topics relate to labels by visually sensing relative distances. *Second*, to support multiple label structures, we further enrich label-dependent topic distribution and derive three variants for single, multiple, and hierarchical labeling, respectively. *Third*, by deterministically supervising observed labels and probabilistically modeling unobserved labels, SemiVN benefits from partially available labels in a semi-supervised manner.

To demonstrate one of SemiVN's use cases, Fig. 1 is a screenshot of an interactive interface of SemiVN's output of the Coronavirus news corpus[1]. Effectively understanding newsstream in terms of their main topics could help in selecting articles of interest to readers efficiently. SemiVN generates topic-word and label-word distributions for semantic interpretability, as seen by the example word clouds. A label is rendered as a black triangle. Right-clicking would reveal its word cloud, which represents a summary of documents. Topics further split a label into sub-concepts. A topic is rendered as a white circle. For instance, the word cloud of the label around the bottom gray area reveals *health and hospital*. In turn, its surrounding topic further focuses on *health situation of Boris Johnson*. To see the content of a document, one can left-click on one of the colored circles (the color reflects the category of the document), revealing the content in a separate window below the scatterplot. The placement of documents on the plot reveals the coherence within each category, as well as the potential semantic relations across labels and topics. For instance, *Economy, business, and finance* category (blue) lies in the center, suggesting that economy is associated with diverse industries and influenced by Coronavirus from many aspects. SemiVN unifies semantic interpretability and visual expressivity and provides a holistic understanding of the corpus.

The joining of visualization and topic modeling, within a semi-supervised framework, lends SemiVN new capabilities not existing in prior models. Pure visualization

[1] https://aylien.com/blog/free-coronavirus-news-dataset



```
Boris Johnson has been admitted to hospital for tests
 after contracting coronavirus, Number 10 has
 confirmed.

The Prime Minister is thought to be seeking medical
 advice as a ‘precautionary measure’.

A Downing Street spokesperson said: ‘On the advice of
 his doctor, the Prime Minister has tonight been
 admitted to hospital for tests.
```

Fig. 1. Semantic visualization of Coronavirus news corpus with 20 topics.

tools, such as the widely used t-SNE [18] does not model topics, and cannot express main topics in the scatterplot. This motivates the development of semantic visualization. Prior works in semantic visualization are mostly unsupervised, except ContraVis [14], which requires full supervision and accommodates only single labels. SemiVN is designed in a semi-supervised manner to leverage a proportion of labeled documents to visualize all and extends beyond single labeling to multiple and hierarchical labeling.

Problem. Let $\mathcal{G} = \{\mathcal{D}, \mathcal{E}, \mathcal{L}\}$ be a document network with labels. $\mathcal{D} = \{\mathbf{d}_i\}_{i=1}^{N}$ is document set. Each document $\mathbf{d} \in \mathbb{R}^{|\mathcal{V}|}$ is a vector in the vocabulary space $\mathcal{V}$. We use *tf-idf* to represent **d**. $\mathcal{E} \subseteq \mathcal{D} \times \mathcal{D}$ contains edges, where $e_{ij} \in \mathcal{E}$ if there is an edge between document i and j. Here we model an undirected network, $e_{ij} = e_{ji}$. We will use *edge* and *link* interchangeably. Document i's neighbors $\mathcal{N}(i)$ are those directly linked to i. As in [28], when no appropriate links are observed in a corpus, one could alternatively induce similarity-based kNN document network based on *tf-idf* cosine similarity. The set $\mathcal{L}$ has observed labels where $\ell_i \subseteq \mathcal{L}$ if we observe document i's label(s) ℓ_i.

Given a partially labeled document network $\mathcal{G}$ as input, the goal is to find visual coordinates for *i)* N documents $\{x_i\}_{i=1}^{N}$, *ii)* T topics $\{\phi_t\}_{t=1}^{T}$, and *iii)* L labels $\{\psi_l\}_{l=1}^{L}$, where the Euclidean distances among coordinates reflect distributions of document-topic, document-label, and label-topic pair.

Contributions. *First*, we introduce coordinate-based label distribution and label-dependent topic distribution, and propose a novel semi-supervised topic model for networked documents that unifies semantic and visual expressivity. *Second*, we extend our model for singly-labeled, multi-labeled and hierarchically-labeled documents. *Third*, our model outperforms baselines quantitatively and qualitatively on public datasets[2].

2 Related Work

Semantic visualization. Incorporating topic modeling into data visualization is referred to as *semantic visualization*. One pioneering model is PLSV [8]. PLANE [15] is the first attempt on semantic visualization of networked documents. However, these models learn coordinates in an unsupervised way, and do not embed labels to reflect corpus hierarchy. While [14] incorporates labels for contrastive visualization, it requires labels for all the documents in the corpus, which precludes the use of unlabeled documents or class prediction. On the contrary, SemiVN unifies networked documents, topics, and labels into the same visualization scatterplot in a semi-supervised way. There are also models [5, 16] visualizing documents, but without the notion of labels as well.

Document network embedding. Previously, topic models for document networks are based on graphical models, e.g., RTM [4] leverages topics of two documents to predict the link. More recent models are based on neural approaches. NRTM [1] extends VAE [10] by introducing a multi-layer perception [2] for link prediction. Adjacent-Encoder [31] models network structure by neighboring document reconstruction. These embed networked documents into topic space only, without any visualization. For the latter, one needs a post-hoc embedding using dimensionality reduction (e.g., t-SNE [18]). In contrast, SemiVN systematically incorporates topic modeling and visualization as a joint approach without the necessity for post-hoc embedding. There are other models that learn node embeddings on attributed graphs [12, 29], but they are not topic nor visualization models. They do not generate topic-word matrix, and the learned embeddings are not topics. Their learning process does not offer visualization.

(Semi-)Supervised topic modeling. Supervised and semi-supervised topic models are those methods that embed both textual content and document labels and produce label-dependent topic distributions. Graphical models include sLDA [20] and DiscLDA [13] for single labeling, LLDA [24] for multi-labeling, and PLDA [25] for partially labeling documents. SemiVAE [11] and MVAE [30] are based on Auto-Encoder, a neural topic model. Similarly, these models do not have an in-built visualization aspect, thus need a post-hoc technique for visual comparison. We also distinguish SemiVN from hierarchical topic modeling, such as nCRP [7], which learns hierarchical topics unsupervisedly. SemiVN's topics are not hierarchical and are semi-supervised.

3 Model Architecture and Analysis

In this section, we describe the technical details of proposed generative approach, whose graphical models are given by Fig. 2. See Table 1 for the summary of notations.

[2] Source code and datasets are available at `https://github.com/cezhang01/semivn`.

Table 1. Summary of notations.

Notation	Description
$\mathcal{G}$	document network
$\mathcal{D}$	document set
$\mathcal{E}$	edge set
$\mathcal{L}$	label set
$\mathcal{V}$	vocabulary
$\mathbf{d}_i$	document i's $tf - idf$ representation in the vocabulary space, $\mathbf{d}_i \in \mathbb{R}^{\lvert\mathcal{V}\rvert}$
$\mathcal{N}(i)$	document i's neighbor set
ℓ_i	document i's observed label(s)
N	number of documents, $N = \lvert\mathcal{D}\rvert$
T	number of topics
L	number of labels
x_i	visualization coordinate of document i
ϕ_t	visualization coordinate of topic t
ψ_l	visualization coordinate of label l
$\hat{\mathbf{y}}_i$	document i's estimated label distribution, $\hat{\mathbf{y}}_i \in \mathbb{R}^L$, or $\hat{\mathbf{y}}_i \in \mathbb{R}^D$ (hierarchical variant only)
$\mathbf{y}_i$	document i's ground-truth label distribution, $\mathbf{y}_i \in \mathbb{R}^L$, or $\mathbf{y}_i \in \mathbb{R}^D$ (hierarchical variant only)
D	depth of the hierarchical softmax tree
H	number of different paths on the tree
h	a path on the tree
M	number of negative samples
$\mathbf{t}_i$	document i's topic distribution, $\mathbf{t}_i \in \mathbb{R}^T$
$\hat{\mathbf{d}}_i$	document i's generated content, $\hat{\mathbf{d}}_i \in \mathbb{R}^{\lvert\mathcal{V}\rvert}$
N_l	number of documents with observed label l
z	dimension of visualization coordinates (2 or 3 in general)

3.1 Coordinate-Based Distribution

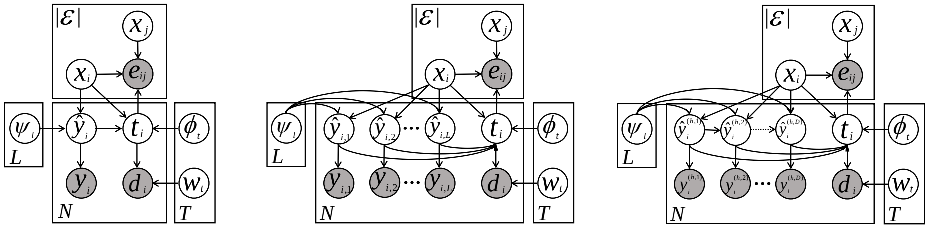
To tightly couple topic modeling and visualization, we devise a model whose parameters are visualization coordinates that give rise to the probability distributions that underlie a topic model. We define coordinate-based label distribution and label-dependent topic distribution, then discuss three modelings: labels $\mathcal{L}$, links $\mathcal{E}$, and text $\mathcal{D}$.

Labels represent main categories of a corpus and summarize a group of topics; topics in turn characterize documents. We preserve corpus structure with a nested approach. First, we introduce a label distribution $p(l|i)$ for document i. The generation of each link e_{ij} can be characterized as follows. For document i, we draw its label $l \sim p(l|i)$, representing i's main category. Its linked neighbor j is then generated based on i and its label l by $j \sim p(j|i,l)$. Formally,

$$p(e_{ij}) = p(j|i)p(i) \propto \sum_l p(j|i,l)p(l|i) \tag{1}$$

where we assume $p(i) = \frac{1}{N}$. Since label is a general description of corpus, and groups a set of topics, given document i and its label assignment, we factorize $p(j|i,l)$ into topic distributions $\sum_t p(j|t)p(t|i,l)$. Eq. 1 can be rewritten as

$$p(e_{ij}) \propto \sum_l p(j|i,l)p(l|i) = \sum_l \sum_t p(j|t)p(t|i,l)p(l|i) = \sum_t p(j|t) \sum_l p(t|i,l)p(l|i). \tag{2}$$



(a) Single-Label SemiVN

(b) Multi-Label SemiVN

(c) Hierarchical-Label SemiVN

Fig. 2. Graphical models of (a) Single-Label SemiVN, (b) Multi-Label SemiVN, and (c) Hierarchical-Label SemiVN.

We interpret Eq. 2 as follows. Each document i is represented by its label distribution $p(l|i)$, wherein each label is decomposed into topic distribution $p(t|i,l)$. In turn, each topic generates neighboring document j by $p(j|t)$. Each link is generated in a nested process, and corpus structure is preserved.

Since we are interested in modeling visualization coordinates, we define label distribution $p(l|i)$ as

$$p(l|i) = \frac{\exp(-\frac{1}{2}||x_i - \psi_l||^2)}{\sum_{l'} \exp(-\frac{1}{2}||x_i - \psi_{l'}||^2)}. \tag{3}$$

This is expressed in terms of the Euclidean distances between a document i's coordinate x_i and those of different labels $\psi_{l'}$. The closer is x_i to a specific ψ_l, the higher is the probability $p(l|i)$, which aligns with the objective of semantic visualization. In turn, for each document and label, we introduce a coordinate-based label-dependent topic distribution $p(t|i,l)$

$$p(t|i,l) = \frac{\exp(-\frac{1}{2}||x_i - \phi_t||^2)\exp(-\frac{1}{2}||\psi_l - \phi_t||^2)}{\sum_{t'} \exp(-\frac{1}{2}||x_i - \phi_{t'}||^2)\exp(-\frac{1}{2}||\psi_l - \phi_{t'}||^2)}. \tag{4}$$

The topic distribution is jointly determined by both document i's coordinate x_i and its label coordinate ψ_l. Document i has a high topic probability $p(t|i,l)$ when it is close to topic ϕ_t, plus ϕ_t is a nearby topic of label ψ_l. Finally, $p(j|t)$ is similarly defined.

$$p(j|t) = \frac{\exp(-\frac{1}{2}||\phi_t - x_j||^2)}{\sum_{j'} \exp(-\frac{1}{2}||\phi_t - x_{j'}||^2)}. \tag{5}$$

So far, we have assumed no label has been observed, and we model such uncertainty in a probabilistic way. Since documents are partially labeled, if we observe document i's label ℓ_i, its deterministic label distribution is $p(\ell_i|i) = 1$ and $p(l \neq \ell_i|i) = 0$. We substitute it into topic distribution ($\sum_l p(t|i,l)p(l|i)$ at Eq. 2), and obtain $p(t|i,\ell_i)$, instead of a summation over all possible labels. We rewrite Eq. 2 below.

$$\begin{aligned} p(e_{ij}|\mathbb{I}(\ell_i)) &\propto \sum_t p(j|t)p(t|i,\mathbb{I}(\ell_i)), \\ \text{where } p(t|i,\mathbb{I}(\ell_i)) &= \begin{cases} \sum_l p(t|i,l)p(l|i) & \text{if } \mathbb{I}(\ell_i)\text{=}\emptyset, \\ p(t|i,\ell_i) & \text{otherwise.} \end{cases} \end{aligned} \tag{6}$$

Here $\mathbb{I}(\ell_i)$ is an indicator on the observation of i's label, $\mathbb{I}(\ell_i) = \ell_i$ if observed, $\mathbb{I}(\ell_i) = \emptyset$ otherwise.

3.2 Label Modeling

Not all corpora share identical label structures. We observe distinct structures that give rise to three variants of SemiVN.

Single-Label. Each document has one label. Still, we observe the labels of only a proportion of documents in the corpus. With coordinate-based label distribution $\hat{\mathbf{y}}_i = p(l|i) = [\hat{y}_{i,1}, \hat{y}_{i,2}, ..., \hat{y}_{i,L}]^T$ estimated by Eq. 3, we maximize the following log-likelihood for document i's observed label ℓ_i.

$$\mathcal{J}_{label} = \log p(\mathbf{y}_i|i) = \sum_{l=1}^{L} y_{i,l} \log \hat{y}_{i,l}. \tag{7}$$

Here $\mathbf{y}_i = [y_{i,1}, y_{i,2}, ..., y_{i,L}]^T$ is the ground-truth label distribution with $y_{i,l=\ell_i} = 1$ and $y_{i,l\neq\ell_i} = 0$.

Multi-Label. SemiVN can be extended to model multi-labeled documents. In this case, document i's observed label set contains more than one label, $|\ell_i| > 1$. Coordinate-based label distribution at Eq. 3 $\hat{\mathbf{y}}_i = [\hat{y}_{i,1}, \hat{y}_{i,2}, ..., \hat{y}_{i,L}]^T$ is no longer softmax. Each single label probability is modified to

$$\hat{y}_{i,l} = \sigma(-\frac{1}{2}||x_i - \psi_l||^2), \quad (l = 1, 2, ..., L), \tag{8}$$

$$\mathcal{J}_{label} = \log p(\mathbf{y}_i|i) = \sum_{l=1}^{L} y_{i,l} \log \hat{y}_{i,l} + (1 - y_{i,l}) \log(1 - \hat{y}_{i,l}). \tag{9}$$

Again $\mathbf{y}_i$ is the ground-truth label distribution. Coordinate-based label-dependent topic distribution Eq. 4 is extended to

$$p(t|i, \ell_i) = \frac{\exp(-\frac{1}{2}||x_i - \phi_t||^2) \prod_{l\in\ell_i} \exp(-\frac{1}{2}||\psi_l - \phi_t||^2)}{\sum_{t'} \exp(-\frac{1}{2}||x_i - \phi_{t'}||^2) \prod_{l\in\ell_i} \exp(-\frac{1}{2}||\psi_l - \phi_{t'}||^2)}. \tag{10}$$

Hierarchical-Label. In contrast to independent labels in the multi-label scenario, hierarchical-label relies on label dependency in a D-level tree (with labels as nodes). Document i's label is thus a path on the tree $\ell_i = \{\ell_{i,d}\}_{d=1}^{D}$. See Fig. 3 for illustration of NET dataset. Motivated by hierarchical softmax [23], we modify coordinate-based label distribution Eq. 3 to $\hat{\mathbf{y}}_i = [\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, ..., \hat{y}_i^{(H)}]^T$, where H is the number of different paths on the tree. The probability of each path is $\hat{y}_i^{(h)} = \prod_{d=1}^{D} \hat{y}_i^{(h,d)}$, and each single-label probability $\hat{y}_i^{(h,d)}$ of d^{th} label on path h is Eq. 8. The log-likelihood function is similar to Eq. 9, except that the summation is in terms of H, rather than L. Finally, its label-dependent topic distribution aligns with Eq. 10.

3.3 Link and Content Modeling

Link modeling. To model all the links in a given network, we maximize the log-likelihood of each observed link, $\log p(e_{ij}|\mathbb{I}(\ell_i))$ at Eq. 6. Directly maximizing this objective is intractable for large networks, we instead maximize its lower bound below.

$$\mathcal{J}_{link} = E_{t\sim q(t|i,j,\mathbb{I}(\ell_i))}[\log p(j|t)] - KL[q(t|i,j,\mathbb{I}(\ell_i))||p(t|i,\mathbb{I}(\ell_i))]. \tag{11}$$

Here $q(t|i,j,\mathbb{I}(l_i))$ is a variational distribution that approximates the true posterior $p(t|i,j,\mathbb{I}(l_i))$, and KL divergence $KL(\cdot||\cdot)$ measures the difference between two distributions. We similarly define variational distribution $q(t|i,j,\mathbb{I}(l_i))$ as

$$q(t|i,j,\mathbb{I}(\ell_i)) = \begin{cases} \sum_l q(t|i,j,l)q(l|i,j) & \text{if } \mathbb{I}(\ell_i)\text{=}\emptyset, \\ q(t|i,j,\ell_i) & \text{otherwise.} \end{cases} \tag{12}$$

We parameterize $q(l|i,j)$ and $q(t|i,j,\ell_i)$ using coordinates.

$$q(l|i,j) = \frac{\exp(-\frac{1}{2}||x_i \oplus x_j - \psi_l||^2)}{\sum_{l'} \exp(-\frac{1}{2}||x_i \oplus x_j - \psi_{l'}||^2)} \tag{13}$$

$$q(t|i,j,l) = \frac{\exp(-\frac{1}{2}||x_i \oplus x_j - \phi_t||^2)\exp(-\frac{1}{2}||\psi_l - \phi_t||^2)}{\sum_{t'} \exp(-\frac{1}{2}||x_i \oplus x_j - \phi_{t'}||^2)\exp(-\frac{1}{2}||\psi_l - \phi_{t'}||^2)}. \tag{14}$$

We use $\oplus$ to denote element-wise average operation. For each link e_{ij}, we utilize Eq. 12 to evaluate topic distribution, and adopt gumbel-softmax reparameterization [9, 19] to sample a topic. Evaluating $\log p(j|t)$ in Eq. 11 is computationally expensive on large networks, since it requires summation over all the documents. Inspired by negative sampling [22], we replace $\log p(j|t)$ with

$$\log\sigma(-\frac{1}{2}||\phi_t - x_j||^2) + \sum_{m=1}^{M} E_{v\sim P_n(v)}[\log(1-\sigma(-\frac{1}{2}||\phi_t - x_v||^2))], \tag{15}$$

$\sigma(x) = \frac{1}{1+\exp(-x)}$ is sigmoid, $P_n(v)$ is a noise distribution over documents, and M is the number of negative samples.

Content modeling. Another important objective is topic modeling by learning topic-word associations. Following previous neural topic models [31, 27], with coordinate-based label-dependent topic distribution $\mathbf{t}_i = p(t|i,\mathbb{I}(\ell_i))$ at Eq. 6, we generate its observed plain text, and parameterize this decoder using a fully connected neural network by $\hat{\mathbf{d}}_i = p(\mathbf{d}_i|\mathbf{t}_i) = \sigma(\mathbf{W}\mathbf{t}_i + \mathbf{b})$. Here $\sigma(x)$ is sigmoid function, $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}|\times T}$ represents topic-word associations, and $\mathbf{b} \in \mathbb{R}^{|\mathcal{V}|}$ is bias. The log-likelihood of the observed textual content $\log p(\mathbf{d}_i|\mathbf{t}_i)$ is

$$\mathcal{J}_{content} = \sum_{w=1}^{|\mathcal{V}|} d_{i,w}\log\hat{d}_{i,w} + (1-d_{i,w})\log(1-\hat{d}_{i,w}). \tag{16}$$

Table 2. Dataset statistics.

Name	#Documents	#Links	Vocabulary	#Labels	Labeling
DS	570	1,336	3,085	9	Single
ML	1,980	5,748	4,431	7	Single
COVID	1,500	6,418	2,226	5	Single
NET	1,278	4,610	6,832	6	Hierarchical
DBLP	14,036	40,269	8,600	4	Multiple

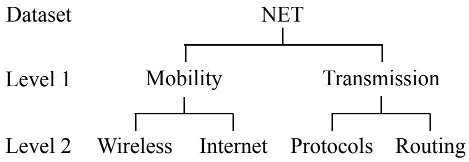


Fig. 3. Label hierarchy of NET.

3.4 The Complete Model

Given a document network $\mathcal{G}$ with links $\mathcal{E}$, document content $\mathcal{D}$, and a proportion of labels $\mathcal{L}$, putting the three components together, we obtain $\mathcal{J} = \mathcal{J}_{link} + \mathcal{J}_{content} + \mathcal{J}_{label}$ as the overall log-likelihood. We intuit that two linked documents are similar if both share many common neighbors. Thus, we add a label smoothness regularizer to objective function, which helps to distinguish different neighbors and encourages strongly connected neighbors to have similar label distributions. Specifically, the regularizer is

$$\mathcal{J}_{reg} = \sum_{e_{ij}} \alpha_{ij} d(p(l|i), p(l|j)), \quad \alpha_{ij} = \frac{|\mathcal{N}(i) \cap \mathcal{N}(j)|}{|\mathcal{N}(i) \cup \mathcal{N}(j)|}. \tag{17}$$

$\mathcal{N}(i)$ is i's neighbor set. As in [26], α_{ij} is a similarity measure based on common neighbors. $d(\cdot,\cdot)$ measures the difference between two distributions. We use KL divergence for single-label and squared difference for multi- and hierarchical-label. Although KL is asymmetric, i.e., $KL(p(l|i), p(l|j)) \neq KL(p(l|j), p(l|i))$, in this paper we model undirected links and consider both e_{ij} and e_{ji}, which removes the effect of asymmetry. Finally, the ultimate loss is (we take negative for log-likelihood for minimization)

$$\mathcal{J} = -\mathcal{J}_{link} - \mathcal{J}_{content} - \mathcal{J}_{label} + \lambda \mathcal{J}_{reg}. \tag{18}$$

λ is a balancing hyperparameter.

Inference. After convergence, in addition to the visualization coordinates, we obtain topic-word association matrix **W** in the content decoder. To infer label-word association, we have $p(\mathbf{d}|l) = \sum_t p(\mathbf{d}|t) \sum_{i\in\mathcal{D}} p(t|i,l)p(i|l)$. Label-dependent topic distribution $p(t|i,l)$ is Eq. 4. $p(i|l) = \frac{1}{N_l}$ if $\ell_i = l$, 0 otherwise. N_l is number of documents with label l. $p(\mathbf{d}|t)$ is content decoder. The keywords (word cloud) of each topic t and label l are those with highest value at $p(\mathbf{d}|t)$ and $p(\mathbf{d}|l)$, respectively. Every topic and label has its own word cloud, but for clarity, we only show some topics and labels. As in previous semantic visualization works, word cloud is not our design, our focus is to extract latent semantics and learn coordinates to visualize documents, topics, and labels. Similarly, as with previous works, including t-SNE, SemiVN is transductive. Our emphasis is using SemiVN to explore existing documents in a corpus for visual and semantic understanding.

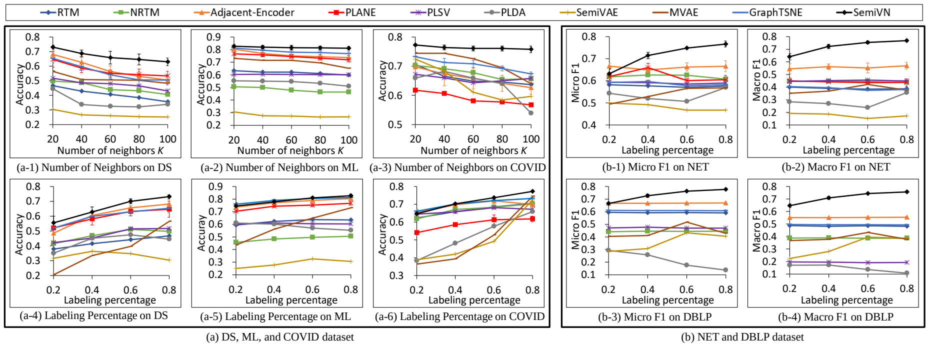
Complexity. Single- and multi-labeling is $\mathcal{O}(\sum_l N_l L z)$, hierarchical-labeling is $\mathcal{O}(\sum_l N_l DHz)$. z is the dimension of visual coordinates (typically 2 or 3). Note that $\sum_l N_l = N$ only if all the documents are labeled, $\sum_l N_l < N$ otherwise. For simplicity, we use F to denote L and DH. Link modeling is $\mathcal{O}(|\mathcal{E}|(zT|\ell_i|_{max} + zM))$. Content modeling is $\mathcal{O}(NT|\mathcal{V}|)$. Putting all three together, we obtain $\mathcal{O}(|\mathcal{E}|(zdT|\ell_i|_{max} + zM) + NT|\mathcal{V}| + \sum_l N_l F z)$. SemiVN converges in one hour on DBLP dataset (see Table 2), while some baseline, PLANE [15], even did not converge in 48 hours. Evaluations were conducted on a machine with Intel Xeon E5-2650v4 2.20 GHz CPU and 256GB RAM.

4 Experiments

Experimental objective is to investigate the quality of visual coordinates. Evaluating visualization is indeed not an easy task. After reviewing many previous visualization works, we summarize some standard experiments, including coordinate classification, link prediction, and topic interpretability as quantitative tasks. In addition, we further conduct user study, involving both static and interactive study.

Datasets. Cora [21] is a public collection of papers with abstract as content and citations as links. Two papers are linked by an undirected link if one cites the other. We extracted three independent datasets, Data Structure (DS), Machine Learning (ML), and Networking (NET). DS and ML contain singly labeled documents, NET is organized into hierarchical labels (Fig. 3). Besides Cora, we created a co-authorship network DBLP. Each author is represented by the aggregation of her publications. Two authors are linked if they have collaboration. If an author publishes at least three papers on one type of conference, we consider her having the corresponding label. Around 11% authors have more than one label. We also created a Coronavirus news corpus. Each article belongs to one category. Since no appropriate links are observed, we generate kNN ($k = 5$) network using $tf - idf$ cosine similarity. Table 2 shows the statistics.

Baselines. We compare to several categories of baselines. *i*) **Topic modeling on networked documents**, including RTM, NRTM, and Adjacent-Encoder. *ii*) **Semantic visualization**. PLSV visualizes documents individually, and PLANE visualizes networked documents. Neither has labels. *iii*) **Semi-supervised topic model**, including PLDA, SemiVAE, and MVAE. In addition, recently there are models for attributed graph embedding. Strictly speaking, they are not topic models, nor baselines. For completeness, we still compare to *iv*) GraphTSNE [17] with GCN [12] and t-SNE as a joint model. Models in category *i*) and *iii*) extract topics only, we pipeline their topics by t-SNE to obtain coordinates. By comparison to these disjoint models, we showcase the advantage of jointly modeling topics and visualization. The comparison to joint models (PLSV, PLANE, GraphTSNE) shows the importance of modeling labels. Each result is obtained by 5 independent runs.



(a-1) Number of Neighbors on DS (a-2) Number of Neighbors on ML (a-3) Number of Neighbors on COVID (b-1) Micro F1 on NET (b-2) Macro F1 on NET

(a-4) Labeling Percentage on DS (a-5) Labeling Percentage on ML (a-6) Labeling Percentage on COVID (b-3) Micro F1 on DBLP (b-4) Macro F1 on DBLP

(a) DS, ML, and COVID dataset (b) NET and DBLP dataset

Fig. 4. Coordinate classification on five datasets.

Hyperparameters are chosen based on validation set. We set 2, 0.01, and 0.01 as Dirichlet prior for RTM, PLANE, and PLSV, respectively. We input texts, adjacency matrix, and labels for MVAE. Other baselines use default settings. For SemiVN, we set the number of negative samples M to 5. Regularizer λ is searched in $[0.1, 0.2, 0.5, 1, 2, 5]$ and set to 1. $tf - idf$ is generated by sklearn (`https://scikit-learn.org`).

4.1 Quantitative Evaluation

Coordinate classification. A good visualization is expected to group coordinates from the same category closely, and separate different categories. For DS, ML, COVID, we adopt K-nearest neighbors as classifier. Our goal is to use labeled coordinates to predict labels for the unlabeled coordinates. We report classification accuracy at $T = 30$ in Fig. 4(a). For clarity, we only report std.dev. of SemiVN and best-performing baselines. We first fix 80% labeling percentage (we further split 10% among them for validation), and vary K for classification. Fig. 4(a-1-3) summarizes the results. Although SemiVN performs similarly with GraphTSNE at $K = 20$ on ML, as K increases, SemiVN stays stable, but GraphTSNE deteriorates its results. This verifies that SemiVN benefits from modeling labels to better separate different groups of coordinates. We then fix $K = 20$ for KNN and vary the percentage of labeled coordinates for training. Fig. 4 (a-4-6) reveals that as labeling increases, most models improve results. SemiVN significantly outperforms baselines on DS. It is competitive with GraphTSNE on ML and COVID, but still outperforms the best topic model, Adjacent-Encoder, showcasing SemiVN's advantage of jointly modeling topics and visualization.

NET and DBLP represent a multi-label classification task, thus we train a one-vs-the-rest logistic regression as a multi-label classifier. We report Micro and Macro F1 scores. We exclude PLANE on DBLP, since it did not converge in 48 hours. Fig. 4(b) presents the results at $T = 30$ when varying labeling percentage. Overall, semi-supervised models tend to improve results with increasing labeling.

Coordinate-based link prediction. Following previous work in semantic visualization [15], we could use coordinates of two documents to predict a link. Following [18], the link probability is $p(e_{ij}) \propto \frac{1}{1+||x_i - x_j||^2}$. For documents with more than three

Table 3. F1 score and AUC of link prediction at $T = 30$ (results are in percentage).

Model	DS		ML		NET		DBLP	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC
RTM	52.4±0.5	80.1±0.7	46.8±0.4	75.0±0.4	47.9±0.5	72.6±0.3	53.5±0.3	81.1±0.2
NRTM	55.7±0.4	83.6±0.8	32.9±0.7	65.6±0.5	47.6±0.7	72.2±0.6	57.6±0.2	59.3±0.1
Adjacent-Encoder	72.8±0.3	94.2±0.4	65.3±0.2	90.3±0.2	67.5±0.1	86.9±0.3	74.0±0.1	**90.3±0.1**
PLANE	75.4±0.3	95.0±0.4	53.3±0.4	90.6±0.4	61.7±0.4	92.6±0.3	-	-
PLSV	59.1±0.6	55.4±1.6	64.6±0.2	50.1±0.2	42.8±0.2	76.6±0.4	45.0±0.5	84.0±0.1
PLDA	59.3±0.9	79.2±1.0	50.6±0.2	79.9±0.4	48.2±0.5	74.7±0.3	53.6±0.4	82.2±0.1
SemiVAE	34.3±0.6	61.1±1.6	38.8±0.3	48.7±1.1	60.6±0.5	55.1±0.6	33.9±0.0	61.8±0.2
MVAE	60.8±0.3	86.6±0.8	42.1±0.8	82.2±0.3	67.3±0.4	71.1±0.3	14.9±0.2	63.0±0.1
GraphTSNE	77.7±0.6	95.7±0.4	65.9±0.4	90.8±0.3	70.2±0.5	90.1±0.2	61.4±0.1	87.9±0.1
SemiVN	**80.5±0.8**	**96.0±0.3**	**79.0±0.9**	**92.9±0.6**	**78.6±0.4**	**92.9±0.1**	**78.3±0.4**	88.2±0.4

Table 4. Topic Coherence NPMI (in percentage) at $T = 30$. GraphTSNE is not a topic model, thus is not included.

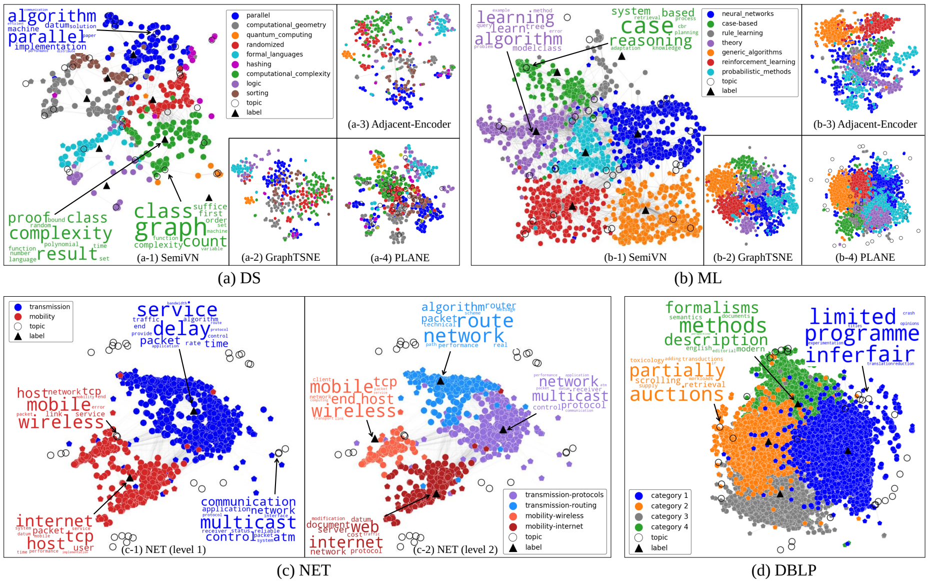
Model	NPMI				
	DS	ML	COVID	NET	DBLP
RTM	8.5±0.4	7.4±0.3	22.8±0.4	14.5±0.7	2.8±0.2
NRTM	6.7±0.3	7.6±0.2	19.5±1.5	12.4±0.2	6.0±0.9
Adjacent-Encoder	5.6±0.5	8.6±0.9	9.9±1.6	11.2±1.2	7.5±1.3
PLANE	8.2±0.1	9.0±0.2	21.1±0.8	14.5±0.6	-
PLSV	8.6±0.2	**10.1±0.3**	**25.9±0.7**	15.4±0.2	6.9±0.6
PLDA	4.0±0.5	2.8±0.3	9.1±0.4	5.0±0.6	4.8±0.2
SemiVAE	3.5±0.8	7.1±1.0	9.1±0.6	8.4±2.1	1.2±1.0
MVAE	4.4±0.6	6.3±0.5	7.9±0.3	11.0±0.9	1.8±0.7
SemiVN (topic)	**9.8±0.3**	**9.9±0.4**	25.0±1.6	**17.6±0.7**	**8.4±0.7**
SemiVN (label)	9.5±0.4	9.7±0.2	24.2±1.1	17.0±0.9 (level 1) 17.2±1.3 (level 2)	8.1±0.2

links, we randomly hold out one. In total 15%-17% links are hidden. We sample the same number of disconnected pairs as negative instances. The remaining network is used for training. Our goal is to predict the held-out links. We report F1 score and AUC in Table 3. As mentioned, we pipeline some baselines with t-SNE. These disjoint models would increment errors from two separate components, thus achieving worse results than SemiVN. SemiVN outperforms PLSV and PLANE, indicating that incorporating labels can indeed help to group related coordinates together and predict links better. We do not report the results on COVID dataset, since this dataset does not explicitly observe links. It contains texts and labels only, and we induce kNN network based on $tf-idf$ cosine similarity. Unlike other datasets where we predict citations or coauthorship, predicting links on COVID does not make any sense in real-word scenarios.

Topic interpretability. We use normalized PMI (NPMI) [3] to evaluate the coherence of top 10 words of each topic. *Google Web 1T 5-gram Version 1* [6] is the external corpus for evaluation. GraphTSNE is not a topic model, thus is excluded. Table 4 shows

Table 5. Classification accuracy (in percentage) of variants at $T = 30$, $K = 20$, 80% labeling.

Model	DS	ML	COVID
SemiVN$-reg$	70.6±1.6	78.7±2.5	70.9±1.0
SemiVN$-\alpha$	72.2±2.0	81.2±2.2	74.3±0.7
SemiVN	**73.1±1.3**	**82.6±1.3**	**77.3±0.3**



(a) DS
(b) ML
(c) NET
(d) DBLP

Fig. 5. Semantic visualization with $T = 30$ topics and 80% labeling (best seen in color).

that SemiVN (topic) generates more coherent words and interpretable topics than others. This supports the importance of labels to improve the quality of topic model.

In addition to those of topics, we also evaluate word coherence of labels. No baseline extracts label-word association. SemiVN (label) is consistently lower than topics', since labels' associated keywords are overly general and capture multiple aspects, resulting in fewer co-occurrences.

Analysis. To evaluate the label smoothness regularizer, we compare to two variants. *i*) SemiVN$-reg$ removes the regularizer. *ii*) SemiVN$-\alpha$ maintains it, but uses the same α_{ij} in Eq. 17, and neighbors are equally important. Table 5 shows that *i*) regularizer is helpful to embed neighbors closely, thus achieves better results; *ii*) modeling neighbors differently is necessary, since disregarding it leads to worse performance.

4.2 Visualization

To sense how SemiVN embeds networked documents, topics, and labels into the same scatterplot, we present visualizations in Fig. 5. See Fig. 1 for COVID. The similarity of topics and labels can be revealed by the relative distance among coordinates. Similar

Table 6. Results of user study.

Question	Adjacent-Encoder	PLANE	GraphTSNE	SemiVN
Q1	10.0%	40.0%	2.5%	**47.5%**
Q2	2.67/5	3.08/5	2.55/5	**4.17**/5
Q3	2.19/5	3.08/5	1.75/5	**4.69**/5

topics and labels tend to group together, distinct ones are separate. Labels and topics split visual space into different semantic subspaces. NET's labels are hierarchical at Fig. 5(c). Label coordinates in Fig. 5(c-1) center within two categories, while in Fig. 5(c-2) the second-tier labels separate into four subspaces. Overall, SemiVN produces clearer separation than baselines.

4.3 User Study

We conduct a user study to test the effectiveness of visualization from human perspectives. We design a survey involving 20 participants who are not authors. The survey comprises 22 questions, of three question types (Q1, Q2, and Q3 below). Each participant is presented with randomly shuffled questions.

- Q1 (MCQ): Given masked plots of 4 anonymized and shuffled models, which best reflects <#labels> clusters?
- Q2 (Rating): Given a colored visualization plot of a model, how good does it separate different categories?
- Q3 (Rating): How related is the clicked article to its surrounding topic and label?

For Q1, we randomly generate visualizations on DS, ML, and COVID. i) We remove topic and label coordinates, and maintain document coordinates only. ii) We use the same color for all the categories. This question looks into the appropriateness of semi-supervision, i.e., if users can identify the correct number of categories from masked plots. For Q2, we consider all five datasets. We go through the same procedure i) as above. Different from ii), we color coordinates based on their own labels. This question tests the separation quality, i.e., if coordinates from different categories are separated. For Q3, we further allow users to interact with visualization. We ask them to select topics of interest based on keywords, then click surrounding article for reading. This investigates if users can use SemiVN to stay informed and select relevant articles in practice, and if the article is visualized at the correct place w.r.t. topics and labels.

We compare SemiVN to three representative baselines: Adjacent-Encoder, PLANE, and GraphTSNE. Q1 contains multiple-choice questions, we report the percentage of participants who favor each model. Q2 and Q3 involve ratings from 1 (terrible) to 5 (excellent), we report the average rating of each model. Table 6 shows that SemiVN outperforms baselines on three types of questions, indicating that users are more satisfied with its ability to visualize documents, topics, and labels. PLANE is the best among baselines, verifying the advantage of modeling topics and visualization jointly.

5 Conclusion

We propose SemiVN, a semi-supervised semantic visualization model that embeds networked documents, topics, and labels into the same visualization scatterplot. Its versatility accommodates different labeling structures: single-label, multi-label, and hierarchical-label variants. Extensive experiments verify the effectiveness of SemiVN in both semantic interpretability and visual expressivity. Future work includes extending SemiVN to inductive scenarios so as to generalize to unseen documents.

Acknowledgments

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Fellowship Programme (Award No. NRF-NRFF2016-07).

References

1. Bai, H., Chen, Z., Lyu, M. R., King, I., Xu, Z.: Neural relational topic models for scientific article analysis. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management pp. 27-36 (2018).
2. Bishop, C. M.: Pattern recognition and Machine learning, Springer (2006).
3. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of GSCL, pp. 31-40 (2009).
4. Chang J., Blei D.: Relational topic models for document networks. In: Artificial Intelligence and Statistics. pp. 81-88 (2009).
5. Choo, J., Lee, C., Reddy, C. K., Park, H.: Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. IEEE transactions on visualization and computer graphics, 19(12), 1992-2001 (2013).
6. Evert, S.: Google web 1t 5-grams made easy (but not for the computer). In: Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop. pp. 32-40 (2010).
7. Blei, D. M., Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B.: Hierarchical topic models and the nested Chinese restaurant process. In: Advances in Neural Information Processing Systems. pp. 17-24 (2004).
8. Iwata, T., Yamada, T., Ueda, N.: Probabilistic latent semantic visualization: topic model for visualizing documents. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 363-371 (2008).
9. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: Proceedings of International Conference on Learning Representations. (2017).
10. Kingma, D. P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. (2013).
11. Kingma, D. P., Rezende, D. J., Mohamed, S., Welling, M.: Semi-supervised learning with deep generative models. In: Proceedings of Advances in Neural Information Processing Systems. pp. 3581-3589 (2014).
12. Kipf, T. N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. (2016).
13. Lacoste-Julien, S., Sha, F., Jordan, M. I.: DiscLDA: Discriminative learning for dimensionality reduction and classification. In: Proceedings of Advances in Neural Information Processing Systems. pp. 897-904. (2009).

14. Le, T., Akoglu, L.: ContraVis: contrastive and visual topic modeling for comparing document collections. In : Proceedings of The World Wide Web Conference. pp. 928-938. (2019).
15. Le, T. M., Lauw, H. W.: Probabilistic latent document network embedding. In: 2014 IEEE International Conference on Data Mining. pp. 270-279. (2014).
16. Lee, H., Kihm, J., Choo, J., Stasko, J., Park, H.: iVisClustering: An interactive visual document clustering via topic modeling. In: Computer graphics forum. Vol. 31, No. 3pt3, pp. 1155-1164. Oxford, UK: Blackwell Publishing Ltd. (2012).
17. Leow, Y. Y., Laurent, T., Bresson, X.: GraphTSNE: a visualization technique for graph-structured data. arXiv preprint arXiv:1904.06915. (2019).
18. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. In: Journal of machine learning research, 9, 2579-2605 (2008).
19. Maddison, C. J., Mnih, A., Teh, Y. W.: The concrete distribution: A continuous relaxation of discrete random variables. In: Proceedings of International Conference on Learning Representations. (2017).
20. Blei, D. M., McAuliffe, J. D.: Supervised topic models. In: Proceedings of Advances in Neural Information Processing Systems. pp. 121-128 (2008).
21. McCallum, A. K., Nigam, K., Rennie, J., Seymore, K.: Automating the construction of internet portals with machine learning. In: Information Retrieval, 3(2), 127-163 (2000).
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems pp. 3111-3119 (2013).
23. Morin, F., Bengio, Y.: Hierarchical probabilistic neural network language model. In: International workshop on artificial intelligence and statistics. pp. 246-252 (2005).
24. Ramage, D., Hall, D., Nallapati, R., Manning, C. D.: Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing. pp. 248-256 (2009).
25. Ramage, D., Manning, C. D., Dumais, S.: Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 457-465 (2011).
26. Rozemberczki, B., Davies, R., Sarkar, R., Sutton, C.: Gemsec: Graph embedding with self clustering. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining. pp. 65-72 (2019).
27. Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488 (2017).
28. Tang, J., Liu, J., Zhang, M., Mei, Q.: Visualizing large-scale and high-dimensional data. In: Proceedings of the 25th international conference on world wide web. pp. 287-297 (2016).
29. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: Proceedings of International Conference on Learning Representations. (2018).
30. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. In: Proceedings of Advances in Neural Information Processing Systems. pp. 5575-5585 (2018).
31. Zhang, C., Lauw, H. W.: Topic modeling on document networks with adjacent-encoder. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34, No. 04, pp. 6737-6745 (2020).