TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections

Minjeong Kim, Kyeongpil Kang, Deokgun Park, Jaegul Choo, and Niklas Elmqvist, *Senior Member, IEEE*

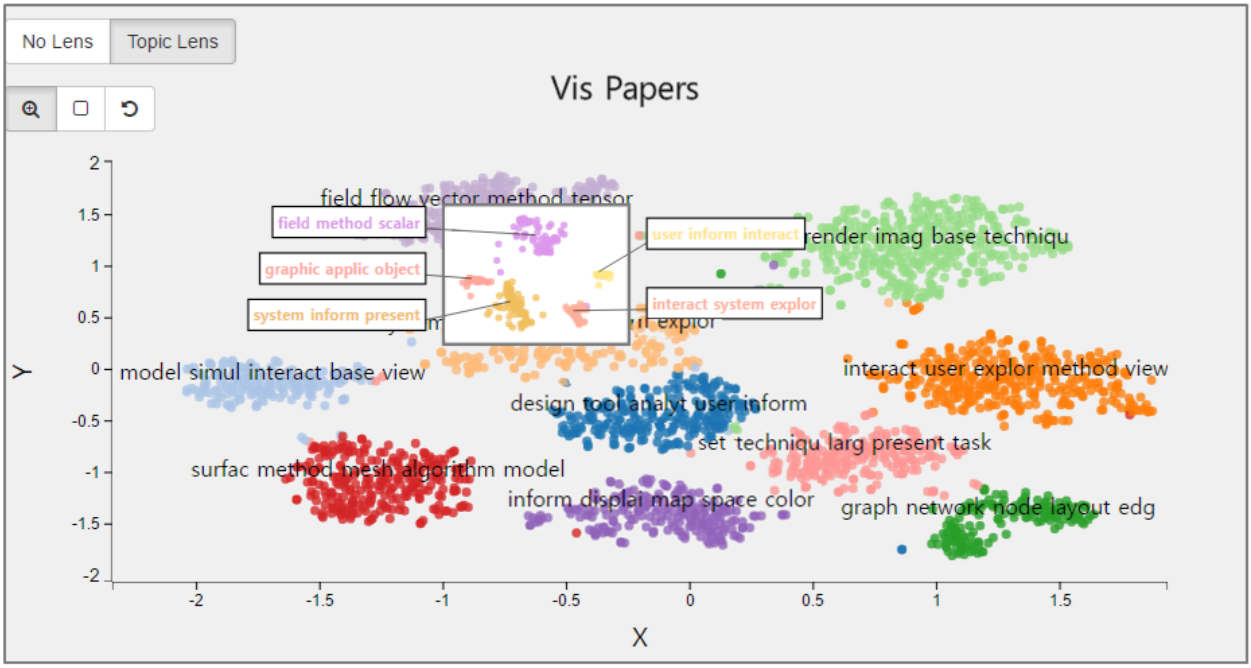


Fig. 1. Overview of our visual analytics system integrated with TopicLens. The system initially performs topic modeling and visualizes documents as a scatterplot where the document coordinates are determined by a 2D embedding method and the topic cluster memberships are color-coded. The representative keywords are shown in the center of each topic cluster. When moving the TopicLens (shown as a small rectangle), we dynamically recompute the topic model and 2D embedding in real time on those documents captured within the lens, revealing their finer-grained topical structure and their visual overview. The representative keywords are visualized just outside of the lens pointing to the center of each topic cluster.

Abstract—Topic modeling, which reveals underlying topics of a document corpus, has been actively adopted in visual analytics for large-scale document collections. However, due to its significant processing time and non-interactive nature, topic modeling has so far not been tightly integrated into a visual analytics workflow. Instead, most such systems are limited to utilizing a fixed, initial set of topics. Motivated by this gap in the literature, we propose a novel interaction technique called TopicLens that allows a user to dynamically explore data through a lens interface where topic modeling and the corresponding 2D embedding are efficiently computed on the fly. To support this interaction in real time while maintaining view consistency, we propose a novel efficient topic modeling method and a semi-supervised 2D embedding algorithm. Our work is based on improving state-of-the-art methods such as nonnegative matrix factorization and t-distributed stochastic neighbor embedding. Furthermore, we have built a web-based visual analytics system integrated with TopicLens. We use this system to measure the performance and the visualization quality of our proposed methods. We provide several scenarios showcasing the capability of TopicLens using real-world datasets.

Index Terms—topic modeling, nonnegative matrix factorization, t-distributed stochastic neighbor embedding, magic lens, text analytics

1 Introduction

How do you automatically summarize all of the articles in a single day from the approximately 1,300 newspapers with regular circulation in the United States? What about 10,000 research articles? A year's worth of press releases from Fortune 500 companies? *Topic modeling* [3, 4] tackles precisely this problem and is one of the most widely used techniques in text mining, natural language processing, and machine learning. The primary goal of topic modeling is to derive a collection of so-called *topics* even from a large-scale document corpus where each topic is represented by a set of coherent keywords that describe a subset of the documents. These topics provide users with a high-level summary of the document corpus without having to read individual documents one by one, and the insights obtained from such a

- *Minjeong Kim and Kyeongpil Kang are with Korea University. E-mail: {mj1642, rudvlf0313}@korea.ac.kr.*
- *Jaegul Choo, the corresponding author, is with Korea University. E-mail: jchoo@korea.ac.kr.*
- *Deokgun Park and Niklas Elmqvist are with University of Maryland in College Park, MD, USA. E-mail: {intuinno, elm}@umd.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

topical summary often lead users to crucial knowledge.

However, while visual analytics systems for large-scale document analysis have certainly adopted topic modeling methods [9, 23, 49], there are two primary issues preventing topic modeling from reaching its full potential when integrated into a visual analytics workflow:

- *Long processing times*: Most topic modeling techniques require significant computation, which is not amenable to real-time usage.
- *Non-interactivity*: Traditional topic modeling techniques do not support an interactive user-guided refinement process.

In practice, this means that most of the current visual analytics systems that include topic modeling can only provide an initial, fixed set of topics. In other words, this precludes interaction between the analyst and the system to refine and extend the topic model for the purpose of improving its quality. Meanwhile, results showed that even state-of-the-art topic models yield an initial output that could get substantial benefit from human refinement [12]. Unfortunately, no established method for interactive topic modeling exists, and the computational demands discussed above make it difficult to introduce such methods.

To address both of these issues at once, we propose TopicLens, a Magic Lens technique [2] for fine-grained interactive topic modeling in a user-specified area of interest. The idea is to let the user selectively refine an overview topic model by moving a lens to a desired area in a 2D scatterplot representing the document corpus. To make this possible, TopicLens builds on two significant technical achievements: (1) a **localized topic modeling** approach based on nonnegative matrix factorization capable of effectively recomputing a topic model for a subset, and (2) a **semi-supervised 2D embedding** based on a t-distributed stochastic embedding that maintains view consistency between the visualization within the lens and the overall visualization. The interactive lens uses *excentric labeling* [19] to render labels at the borders of the lens to avoid obscuring its local contents. In our web-based implementation of TopicLens in a visual analytics system for document analysis (Fig. 1), we demonstrate the computational performance as well as the interactive capabilities of our proposed contributions. Furthermore, we also showcase the TopicLens approach in action for several real-world datasets.

The remainder of this paper is structured as follows: In Section 2, we present the related work on interactive topic modeling for visual analytics. We then present our proposed TopicLens technique in Section 3. Next, in Section 4 and Section 5, we describe our experiments and usage scenarios, respectively. The strengths and the weaknesses of our approach are discussed in Section 6. Finally, we end with our conclusions and visions for future work in Section 7.

2 Related Work

In this section, we discuss the related work from two specific perspectives: interactive lenses and topic modeling for visual analytics.

2.1 Interactive Lenses

Interactive lenses controlled by the user are widely used in general interfaces to reveal hidden or detailed information. Lens techniques are defined as focus+context techniques since they operate on the visual representation itself; the alternatives are overview+detail techniques, which use a separate window to show an alternate view of a visual representation. Several surveys exist on these practices [13, 34].

A magnifying glass is the canonical example of a focus+context lens, where the magnified area (the focus) is naturally integrated into the surrounding visual representation (the context). Appert et al. [1] studied several high-precision variations of such magnification lenses. However, normal magnifying lenses have a drawback in that the lens itself occludes parts of the underlying visual representation. Both the DragMag [48] and PolyZoom [27] techniques try to solve this problem by placing the lens focus outside the viewport, but this then introduces a spatial separation between the focus and the context.

To remedy this problem, distortion-based techniques deform the visual representation to seamlessly integrate the focus into the context. The first such technique, the fisheye view [21, 47], achieves this using nonlinear distortion. The Table Lens method [38] highlights specific rows or columns of a table while maintaining the overall structure by distorting the table layout. However, although distortion yields seamless views, the nonlinear deformation causes visual instability and makes it difficult for users to build a mental model of the space [39,40].

Alternatives that overcome these limitations have been proposed recently. Carpendale's elastic representations [6] generalize the shapes and distortion parameters of interactive lenses. Sigma Lenses [36] reduce the effects of distortion by transitioning the view over space and time, and by varying the transparency. Magic Lenses [2] are special lenses that replace the visual representation of the object inside the lens instead of magnifying its contents.

2.2 Lenses in Visualization

While the previous section discussed the general use of interactive lenses in human-computer interaction, lens techniques are particularly useful for visual analytics as well. From the perspective of Shneiderman's visual information-seeking mantra, *"Overview First, Zoom and Filter, and Details on Demand"* [41], the lens is a powerful interaction technique for visualization and visual analytics that mainly serves the "zoom and filter" step. Tominski et al. [44] presented a general survey on this topic; here, we discuss the work directly relevant to our contribution.

Visualizations are often characterized by complex visual representations, and the most powerful lens techniques are those that creatively combine with filtering or Magic Lens approaches, where the underlying visual representation is changed or simplified. For example, excentric labeling [19] makes it possible to selectively label data items through a user-controlled lens. Similarly, the EdgeLens method [50] alleviates the edge congestion problem in large-scale, complex graph data by locally reducing edge crossings and bending edges inside the region of interest. Tominski et al. [43] integrated several lenses based on both distortion and non-distortion to easily expand and collapse the vertices of a tree or a graph. The Color Lens method [18] is a Magic Lens technique that locally changes the color scale inside its focus to show higher data resolution. Finally, the VectorLens method [17] allows brushing and filtering data-mapped curves based on their angle or direction.

Even with these long-standing research efforts on using interactive lenses in visualization, to our knowledge, no previous studies have tried to integrate it with computationally intensive analytic techniques such as topic modeling. In this respect, our TopicLens method is one of the first approaches in this novel direction of research, which achieves a tight integration of analytic components with visual analytics.

2.3 General Topic Modeling

Topic modeling is a form of text mining where patterns and themes are identified in a document corpus using statistical methods. The prominence of these methods has become increasingly important as the amount of document data continues to grow exponentially. Several different methods exist; we review the important ones here.

Latent semantic indexing (LSI) [16] can be viewed as one of the earliest topic modeling methods based on applying a well-known matrix factorization technique called singular value decomposition [22] on a term-document matrix. However, the fact that LSI allows both positive and negative weight values of keywords in a topic makes it difficult for a user to interpret the results. In response, probabilistic topic modeling methods have been proposed, where a topic and a document are modeled as (nonnegative) probability distributions over keywords and topics, respectively [3]. Probabilistic latent semantic indexing (pLSI) [25] and latent Dirichlet allocation (LDA) [4] are two popular methods in this category, and, in particular, LDA is currently one of the most widely used topic modeling methods. However, a disadvantage of these methods is their high-performance requirements.

More recently, nonnegative matrix factorization (NMF) [31] has been proposed as an alternative topic modeling approach in document analysis [29]. NMF basically performs matrix factorization with nonnegative constraints, whose outputs are always nonnegative just like

those of probabilistic methods. Thus, it does not suffer from the interpretation difficulties of LSI. Furthermore, compared to probabilistic methods, NMF has shown its advantages in terms of running time and algorithmic consistency [9].

2.4 Topic Modeling in Visual Analytics

The main purpose of using topic modeling in topic modeling is generally to help users interactively explore document data and extract their relationships through topic summaries for the entire corpus. Focusing on the analysis of the topic modeling output itself, Iwata at al. [26] analyzed the topic modeling outputs from pLSI and LDA in a static scatterplot generated by 2D embedding. Termite [11] provided an interactive analysis of the quality of extracted topics via a matrix view that visualizes the term-by-topic association. Chaney et al. [7] developed an interactive system that allows a user to explore different topics along with their associated keywords and documents.

Topic modeling has also been integrated with more sophisticated visual analytics methods for particular data analysis tasks, such as time, connections, and embeddings. TIARA [49] is one such system, and it shows the topical evolution of streaming document data. To this end, TIARA adopts a ThemeRiver style of visualization [24]; many other visual analytics systems have since improved upon this type of visualization [14, 33]. Similarly, FacetAtlas [5] utilizes graph layout-based visualization to aid users in exploring the multi-faceted relationships between topic clusters. Finally, TopicPanorama [35] reveals the connections between topics from multiple heterogeneous document corpora.

While all of these systems are interactive, they tend to use static topic modeling results rather than allow for interactively steering the topic modeling process. The reason is primarily because of the high-performance requirements of topic modeling, which makes it impractical for real-time integration. There exist a few exceptions, however. TopicNets [23] iteratively recomputes topic modeling results on a dynamically changing subset of documents that a user navigates through. iVisClustering [32] provides an interaction capability that iteratively recomputes topic models on a document subset where noisy documents can be excluded. Finally, UTOPIAN [9] offers several nontrivial interaction capabilities by directly steering the topic modeling, e.g., changing the keyword weights of a topic, splitting and merging topics, and creating a new topic based on a seed keyword or document.

In most of the above-described studies, however, the highly dynamic interactions with topic modeling that require efficient, real-time computations of topic modeling have not been explored. In this sense, our TopicLens technique for real-time topic modeling and stable 2D word embedding opens up a new level of interaction. With topic modeling, users can receive finer-grained topical information on a highly dynamic subset of documents that they select themselves.

3 TopicLens: Localized Interactive Topic Modeling

TopicLens is a novel interaction technique that performs topic modeling dynamically on a document subset of interest that a user selects. The technique allows a user to flexibly drill down to a fine-grained topic information about the subset. In this section, we start with an overview of our visual analytics system in which TopicLens is integrated. Then, we present our novel topic modeling and 2D embedding algorithms that accomplish a real-time lens interface while maintaining the consistency between global (outside the lens) and local (inside the lens) context. Finally, we discuss how we further improved real-time interactivity using the idea of progressive visual analytics.

3.1 System Overview

We built a sophisticated web-based visual analytics system centered around our TopicLens technique (Fig. 1). Initially, the main view shows the overview of an entire dataset as a scatterplot by applying topic modeling and 2D embedding. The system also color-codes each document in terms of its most closely related topic cluster. At the center of each topic cluster in the scatterplot, the most representative keywords are displayed so that a user can obtain the topical summary

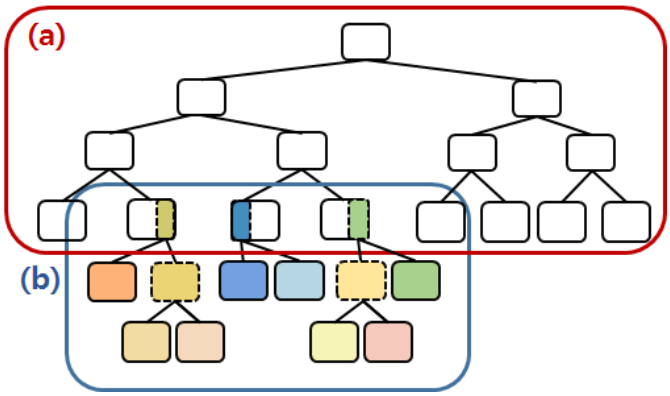


Fig. 2. An initial binary topic tree built by H-NMF (a) and another binary topic tree dynamically generated by our DH-NMF (b) for a document subset captured within the lens.

of the entire data from the scatterplot. By default, the number of topics used to generate an initial topic modeling result is set to 10.

Topic Modeling. For the initial topic modeling, we use a recently proposed hierarchical topic modeling based on recursive rank-2 nonnegative matrix factorization (H-NMF) [30]. By default, the initial number of topics is set to 10. As shown in Fig. 2(a), H-NMF constructs a binary tree of topic clusters given an entire document corpus, where each leaf node corresponds to a single topic that contains an associated document subset. The reason for using this method is twofold. First, it yields a significant improvement in computational time over standard nonnegative matrix factorization [28] and other topic modeling methods such as latent Dirichlet allocation [4]. Furthermore, an initial, hierarchical topic structure makes it efficient to dynamically split/merge the corresponding leaf nodes that contain those documents captured in the lens (Fig. 2(b)). More details about our algorithm for this process will be described in the next section.

Two-Dimensional Embedding. To generate the 2D scatterplot of documents, we use a supervised version [9] of t-distributed stochastic neighbor embedding (t-SNE) [46]. In general, document data are loosely clustered, and, thus, their 2D embedding results tend to overlap with each other among topics, which prevents a user from properly obtaining a high-level topical overview. To avoid this problem, the supervised t-SNE changes the input pairwise distance matrix in a way that those distances within the same topic cluster become closer by a particular factor, while those distances across different topic clusters become farther by another particular factor. In this manner, the topic clusters become clearly separated in a scatterplot, as seen in Fig. 1.

TopicLens. Given the initial scatterplot, a user can dynamically perform the interactions provided by TopicLens by simply dragging a lens onto the clusters or the document subset she/he intends to analyze. Once a user places the lens at a particular place, TopicLens automatically computes the topic modeling on the data captured inside the lens and generates a new scatterplot based on it (Fig. 1). In addition, TopicLens uses excentric labeling [19] to show the representative keyword labels of each topic at the left or the right borders of the lens to prevent these labels from obscuring the visualization inside the lens.

3.2 Dynamic Hierarchical Rank-2 Nonnegative Matrix Factorization

The capability of real-time computation of topic modeling is the key requirement for achieving the highly dynamic interactions provided by TopicLens. To this end, we propose a novel topic modeling approach called dynamic hierarchical rank-2 nonnegative matrix factorization (DH-NMF). Our method is built based on a recently proposed hierarchical rank-2 nonnegative matrix factorization (H-NMF) [30], which has shown superior efficiency and output quality in real-world applications.

Standard NMF. To begin with, standard NMF performs topic modeling as follows. Suppose that we are given a document dataset represented as a term-document matrix $X \in \mathbb{R}_{+}^{m \times n}$, which contains n documents composed of m keywords. Given the number of topics $k \ll \min(m,n)$, NMF computes the low-rank approximation of X, i.e.,

$$\min_{W,H \geqq 0} \| X - WH \|_F^2, \tag{1}$$

where $W \in \mathbb{R}_{+}^{m \times k}$ and $H \in \mathbb{R}_{+}^{k \times n}$. In the two output matrices W and H, W represents a set of k topics, where each column, corresponding to each topic, is described as a weighted combination of m keywords. In this case, as the value of an element gets larger in a particular topic, the corresponding keyword is considered to be more relevant to the topic. On the other hand, H represents a set of n documents, where each column, corresponding to each document, is described as a weighted combination of k topics. In a clustering setting, the topic associated with the largest value in each column of H determines the topic cluster membership of the corresponding document.

Hierarchical NMF. Basically, H-NMF [30] performs the hierarchical clustering of a given document by constructing a binary topic hierarchy, as shown in Fig. 2. In detail, H-NMF successively performs the low-rank approximation with $k = 2$ in Eq. 1 for those documents contained in each node of the hierarchy, which then splits them into two groups corresponding to two child nodes, respectively. When one wants to obtain k topics, such a recursive splitting process of H-NMF continues until the total number of leaf nodes in the binary topic tree becomes k. The criterion for determining which node to split is based on the score estimated by the modified normalized discounted cumulative gain (mNDCG), which measures how different the two newly created topics (corresponding to two child nodes) are from the topic of their parent node.

Exploiting the special algorithmic characteristics of NMF with $k = 2$, H-NMF runs significantly faster than the standard NMF in generating the same number of topics.

Dynamic Hierarchical NMF. The constructed topic hierarchy generated from H-NMF has important advantages in user-driven topic modeling. It can flexibly give a topical overview at a different level, depending on user needs. Furthermore, it is suitable for a user to locally change the hierarchy so that a user can drill down to the document subset of interest.

Further improving H-NMF for TopicLens, we propose dynamic H-NMF (DH-NMF), which can serve as a real-time topic modeling approach for a dynamically changing document set. Our main idea is to utilize an initially built topic hierarchy structure from H-NMF. Suppose that those documents captured in the lens belong to k_i different topic nodes in the initial topic hierarchy and that we want to obtain k_s topics in total, where $k_s \geq k_i$. Fig. 2(b), for example, shows the case where $k_i = 3$ and $k_s = 8$. In this situation, DH-NMF works as follows:

1. We update the m-dimensional topic vector of each of these nodes as the centroids of the bag-of-words vectors of captured documents in each node. These updated k_i topic vectors reflect only those documents captured in the lens while maintaining the initial topic hierarchy.

2. Starting with the k_i updated topics along with the k_i corresponding nodes as multiple root nodes, we continue splitting them based on the mNDCG criterion until we obtain k_s leaf nodes.

DH-NMF has the two main advantages for TopicLens: computational time and topic consistency. First, the extra computational time saving compared to H-NMF is obtained because DH-NMF starts with k_i multiple root nodes instead of a single root node. In this manner, to obtain k_s leaf nodes in total, DH-NMF needs to perform only $(k_s - k_i)$ number of binary splitting operations, each of which corresponds to a single run of rank-2 NMF, while H-NMF needs to perform $(k_s - 1)$ number of them. For example, Fig. 2(b) shows $(8-3)$ binary splitting operations to obtain eight leaf nodes or topics for the documents captured in the lens. Furthermore, it usually takes much more time to perform rank-2 NMF for those nodes near the root level of hierarchy since they involve more documents than those near the leaf level. Because of this fact, DH-NMF is significantly faster than H-NMF, which makes it suitable for TopicLens.

Second, another advantage of DH-NMF is that it maintains the topic consistency between the views both outside and inside the lens. That is, DH-NMF does not merge the initial topics at all, but it only splits them, revealing the subtopics of the original topics existing in the initial topic hierarchy. In this manner, TopicLens helps a user maintain the global context when exploring the new subtopics shown in the lens.

3.3 Guided Approximate t-Distributed Stochastic Neighbor Embedding

For the homogeneous visualization with the main scatterplot view, TopicLens visualizes the new topic modeling results in a scatterplot form via 2D embedding. To this end, we propose a novel 2D embedding algorithm based on one of the state-of-the-art techniques, t-distributed stochastic neighbor embedding (t-SNE) [46], where we achieved (1) real-time computational efficiency as well as (2) consistency with the global view.

t-SNE. Basically, t-SNE is a dimensionality reduction approach that embeds the original high-dimensional data into a low-dimensional (typically 2D) space so that their original pairwise relationships can be maximally preserved. The overall process of t-SNE can be summarized as follows:

1. Given a set of the original m-dimensional vectors, x_i's, of n data items for $i = 1, \cdots, n$, where m denotes the vocabulary size, t-SNE computes the pairwise (m-dimensional) Euclidean distance matrix $D^p \in \mathbb{R}^{n \times n}$, which is then converted into a joint probability matrix $P \in \mathbb{R}^{n \times n}$ so that a bigger pairwise distance value can be converted to a lower probability. Specifically, by adopting a Gaussian distribution for this conversion, we can compute the (i, j)-th component p_{ij} of P as

$$p_{ij} = \frac{\exp(- \| x_i - x_j \|^2 / 2\sigma^2)}{\sum_{k \neq l} \exp(- \| x_k - x_l \|^2 / 2\sigma^2)}. \tag{2}$$

2. t-SNE randomly initializes the 2D embeddings, y_i's, of n data items for $i = 1, \cdots, n$, and computes their Euclidean distance matrix $D^Q \in \mathbb{R}^{n \times n}$. This matrix D^Q is then converted into a joint probability matrix $Q \in \mathbb{R}^{n \times n}$ by adopting a Student's t-distribution. Specifically, the (i, j)-th component q_{ij} of Q is computed as

$$q_{ij} = \frac{(1 + \| y_i - y_j \|^2)^{-1}}{\sum_{k \neq l} (1 + \| y_k - y_l \|^2)^{-1}}.$$

3. t-SNE iteratively updates each y_i of n data items based on the gradient descent with respect to the objective function as the Kullback–Leibler divergence between P and Q, i.e.,

$$C = KL(P \| Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

   Intuitively, this process is similar to the traditional force-directed layout. Given a particular y_i, each of the remaining data items, y_j's, for $j = 1, \cdots, n$ and $j \neq i$, works as either an attractive or a repulsive force, depending on whether the original probability p_{ij} is bigger than the current probability q_{ij} or not.

Approximate t-SNE. For the sake of the real-time performance of 2D embedding in TopicLens, we propose a straightforward approach for accelerating t-SNE, which we call approximate t-SNE. To realize this, we adopted a sampling approach that has been applied in other well-known dimensionality reduction techniques such as multidimensional scaling (MDS) [15, 51]. Our approximate t-SNE utilizes

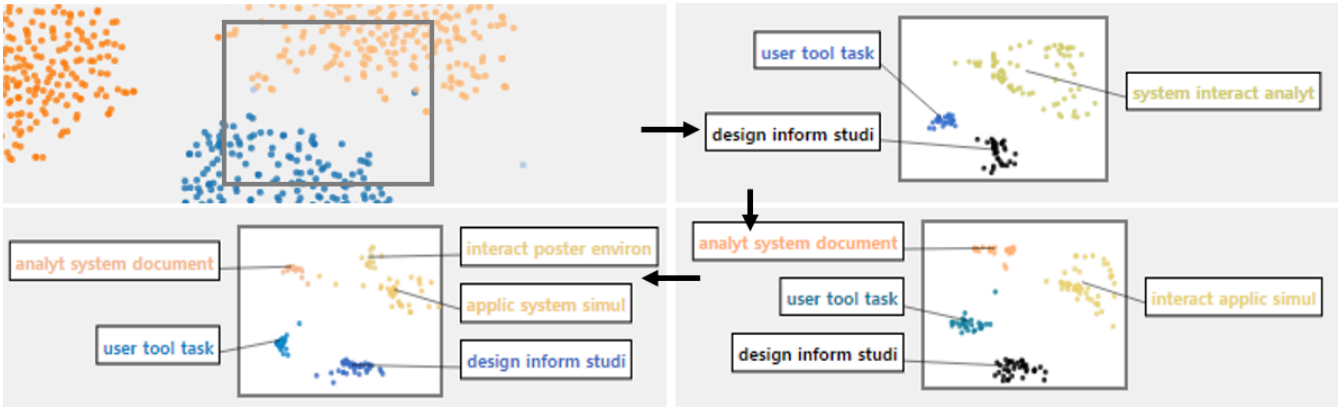


Fig. 3. Progressive visualization of topic modeling results in TopicLens. Once the lens is placed, TopicLens progressively visualizes the intermediate topic modeling outputs in real time, as our topic modeling method, DH-NMF, gradually generates additional topics over time.

only a fraction of the data items to compute the probabilities in Q in Step 2 and the gradient in Step 3 in t-SNE.

In detail, given a landmark ratio r $(0 < r < 1)$, we first sample rn data points as landmark points, the set of which is denoted as $\mathscr{L}$. Then, in Step 2, we compute q_{ij} by using only the landmark points, i.e.,

$$q_{ij} = \frac{(1+\parallel y_i - y_j \parallel^2)^{-1}}{\sum_{k \neq l}(1+\parallel y_k - y_l \parallel^2)^{-1}} \text{ for } \forall j \in \mathscr{L}.$$

Next, in Step 3, we update each y_i with the gradient descent with respect to the new objective function, which involves only the landmark points, i.e.,

$$C = KL(P \parallel Q) = \sum_i \sum_{l \neq i, l \in \mathscr{L}} p_{il} \log \frac{p_{il}}{q_{il}}.$$

In this equation, since the gradient involves only rn iterations instead of n iterations when updating y_i, we can obtain the computational saving by a factor of r. In this manner, our approximate t-SNE achieves a better computational complexity of $O(rn^2)$, compared to the original computational complexity of t-SNE, $O(n^2)$. However, as r reduces towards zero, the approximation becomes more drastic while computational time decreases. In Section 4.2.2, we will discuss the effect of r and our choice for TopicLens that gives the optimal trade-off between the approximation error and the computational time saving.

Guided t-SNE. The second technical novelty we create for improving t-SNE is what we call guided t-SNE. The main purpose of guided t-SNE is to make the 2D embedding in the lens consistent with the global 2D scatterplot outside the lens. For instance, if particular data points were originally placed in the top-left corner in the area captured by the lens, then it would be ideal to place them roughly in the same region in the new scatter plot inside the lens. This way, TopicLens can provide a 2D embedding consistent with the global view.

To achieve this goal, we introduce the notion of anchor points and utilize them as additional data points in our guided t-SNE algorithm. To be specific, given a subset of x_i's captured in the lens, let us denote their initial 2D coordinates in the global scatterplot generated by the initial t-SNE as y_i^G's. Once the new topic modeling result, say, k_s topics, is computed for the subset, then for each of the k_s topics, we compute the centroid by taking the average of y_i^G's for those documents sharing the same topic cluster membership, which results in k_s centroids, c_i's. Next, for each x_i, we set the ideal distance between x_i and the centroid corresponding to its topic cluster as a particular value d_c. This additional distance value per data item is converted to a probability using Eq. (2), and the joint probability matrix P now has an additional row and a column, i.e., $P \in \mathbb{R}^{(n+1)\times(n+1)}$, where the last row and the column contain the probability between each point and its corresponding topic cluster centroid. Next, we include such topic cluster centroids c_i's as virtual points in the 2D embedding space and also add the gradients for each y_i incurred by c_i's when performing the iterative optimization. In addition, during the optimization, these virtual points, c_i's, remain unchanged in the 2D embedding instead of being updated by other points.

Intuitively, this process can be viewed as a weakly constrained or guided process of t-SNE, which prevents significant changes between the previous 2D embedding results and the newly computed ones. However, our method performs such a constraint process at a topic cluster level instead of at an individual data item level so that the topical consistency can be maintained. Furthermore, the parameter d_c in our guided t-SNE determines how strongly y_i's should be tied with their corresponding topic cluster centroid. A smaller d_c not only will make the 2D embedding result more compactly clustered, but also will make it more consistent with the previous 2D embedding. Finally, guided t-SNE exposes users to an overall context even inside the lens and prevents them from being detached from the global context while dynamically exploring the subset of documents via TopicLens.

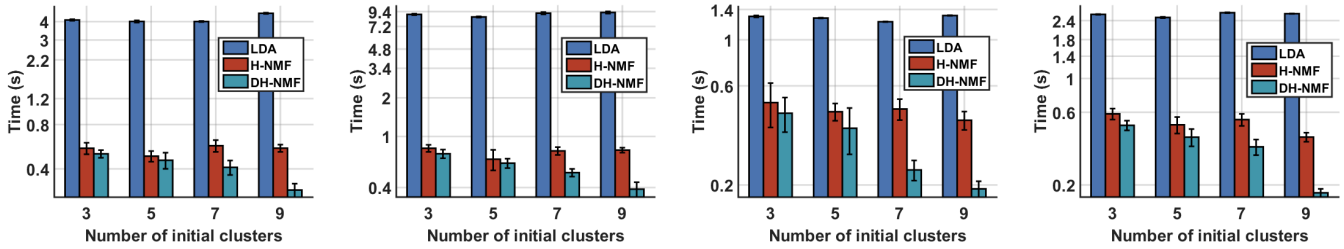
In TopicLens, we use approximate guided t-SNE that combines both of these approaches we proposed above so that we can achieve real-time response and consistency with a global view at the same time.

3.4 Progressive Visualization with Topic Modeling

Highly responsive real-time visualization is the key requirement for TopicLens. Even though our proposed approaches for topic modeling and 2D embedding bring significant efficiency gain, a user may still want to check the results immediately even before the entire computations complete. To address this issue, there exist previous studies [8, 10, 20, 42] that attempted to support real-time interaction with intermediate results while the algorithm proceeds in the background.

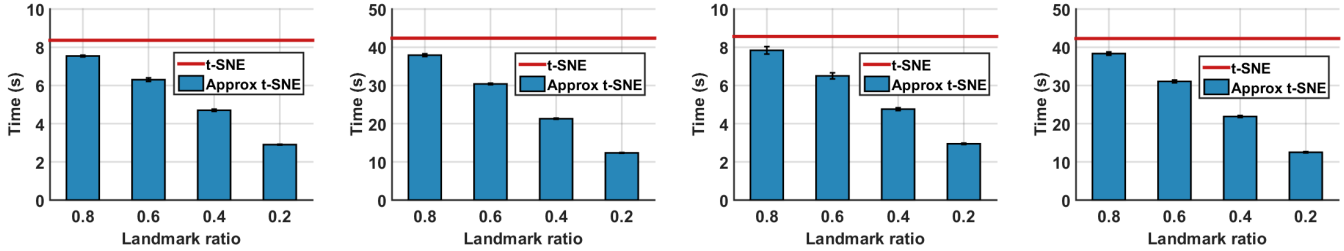
We leverage this idea in TopicLens to visualize the progressive outputs from topic modeling even before they are fully generated. As described in Section 3.2, DH-NMF keeps growing from the initial topic hierarchy tree until we obtain k_s leaf nodes. While DH-NMF generates one topic at a time by splitting a node in the hierarchy during this process, we progressively visualize the topic modeling output at each step in real time, as shown in Fig. 3. In addition, we initiate this progressive visualization when we have just initial topic results with their updated centroid vectors, even before splitting any nodes.

Furthermore, our progressive visualization is not just limited to revealing the progress of DH-NMF, but it also continuously visualizes the progressive outputs of our guided approximate t-SNE. In this manner, we truly achieve the real-time responsiveness from both topic modeling and 2D embedding algorithms.



(a) NYT dataset ($n = 300$). (b) NYT dataset ($n = 600$). (c) VisPub dataset ($n = 300$). (d) VisPub dataset ($n = 600$).

Fig. 4. Comparisons of the computing times between LDA, H-LDA, and DH-LDA, depending on the number of initial topic clusters. DH-NMF shows the fastest computing times, which makes TopicLens efficient. In addition, the performance margin becomes larger as the number of initial topic clusters increases.



(a) NYT dataset ($n = 400$). (b) NYT dataset ($n = 800$). (c) VisPub dataset ($n = 400$). (d) VisPub dataset ($n = 800$).

Fig. 5. Comparisons of the computing times between standard t-SNE and approximate t-SNE. The red line corresponds to the computing time of the original t-SNE, while the bar graphs represent those of approximate t-SNE. As the landmark ratio r gets smaller, approximate t-SNE shows better performances than standard t-SNE.

3.5 Implementation Details

We built our TopicLens-enabled visual analytics system as a web-based application developed with D3[1] and AngularJS.[2] For guided approximate t-SNE, we implemented its algorithm in JavaScript based on the original t-SNE code.[3] For DH-NMF, we implemented it based on the original H-NMF code[4] in MATLAB and it communicates with the client side using the Python Flask micro web framework[5] and the MATLAB Engine for Python.[6] For the progressive visualization discussed in the previous section, we utilized socket communications using flask-SocketIO[7] for the server side and [8] for the client side.

4 Experiments

In this section, we present algorithmic evaluations to examine the quality of our proposed methods from two different perspectives: (1) computing times and (2) consistency with the global view.

4.1 Datasets

We chose two datasets for our experiment: (1) New York Times articles (**NYT**) and (2) academic papers published in the areas of visualization (**VisPub**).

For the NYT dataset, we crawled it from the New York Times website.[9] We collected news articles containing the search query "North Korea" that were published from 2011 to 2015. Since these articles were generated from a specific topic—North Korea—we excluded frequently appearing but less meaningful words such as "north," "south,', "Korea," "Kim," etc. The NYT dataset contained 3463 articles consisting of 22,496 words in total.

The VisPub dataset is a collection of academic papers published in the IEEE Visualization Conference from 1990 to 2014. This collection includes various structured and unstructured fields such as abstract, author, body, and title. Like in the NYT dataset, we also excluded the dominant words in this domain, such as "visualization," "visual," "analysis," etc., to obtain a comprehensive set of topics. Finally, the VisPub dataset contained 2592 documents composed of 12,788 words.

4.2 Computing Times

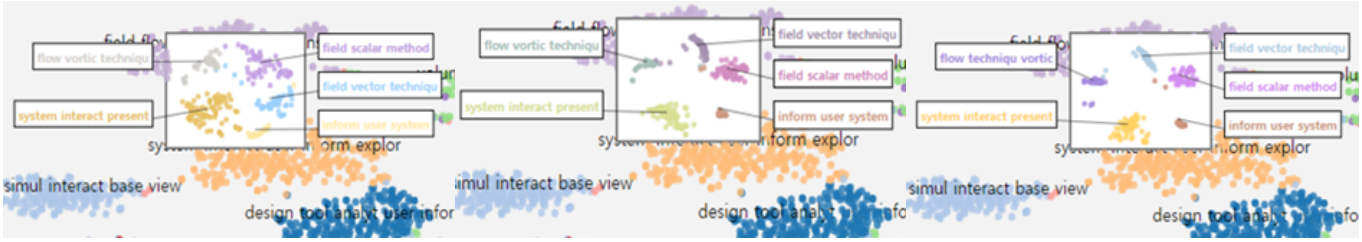
Here, we present two experimental results that show the advantage of DH-NMF and approximate t-SNE algorithms in achieving real-time response in TopicLens.

4.2.1 Dynamic Hierarchical Rank-2 NMF

In this experiment, we compared the computing times between LDA, H-NMF, and DH-NMF when generating k_s number of sub-clusters from a given number of initial clusters. These methods are summarized as follows:

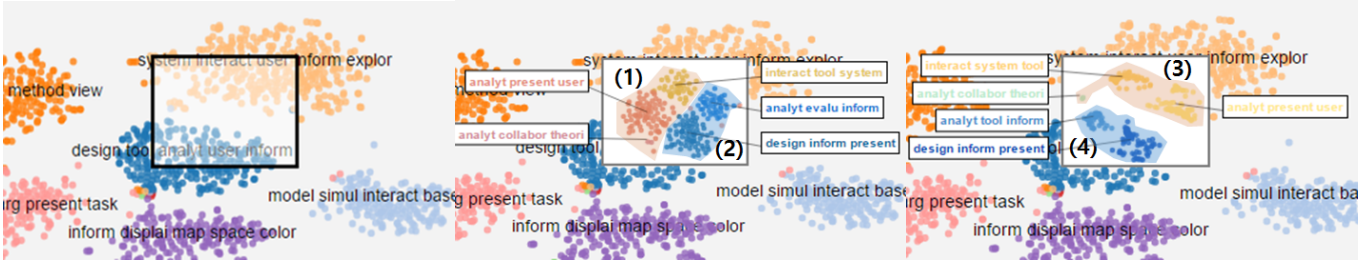
- LDA: Latent Dirichlet allocation, a generative probabilistic topic modeling method [4] based on the Gibbs sampling method. [37] We used the code provided by MATLAB Topic Modeling Toolbox 1.4.[10] We used the default model parameters, and the total number of iterations was set to 1000.
- H-NMF: Hierarchical rank-2 NMF [30], a hierarchical clustering and topic modeling method, which is based on rank-2 NMF. We used the code obtained from the original author's website.[11]

[1] https://d3js.org/
[2] https://angularjs.org/
[3] https://github.com/karpathy/tsnejs
[4] http://math.ucla.edu/ dakuang/software/rank2_safe.zip
[5] http://flask.pocoo.org/
[6] http://mathworks.com/help/matlab/matlab-engine-for-python.html
[7] https://flask-socketio.readthedocs.org/en/latest/
[8] http://socket.io/
[9] http://www.nytimes.com/
[10] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm
[11] http://math.ucla.edu/ dakuang/software/rank2_safe.zip



(a) $r = 0.2$ (b) $r = 0.6$ (c) $r = 1$

Fig. 6. Visualization examples of approximate t-SNE in TopicLens as the landmark ratio r changes. Even with a small value of r, e.g., 0.2, TopicLens shows the overall topical structure relatively well.



(a) The area where TopicLens is applied. (b) Standard t-SNE. (c) Guided t-SNE.

Fig. 7. Effects of guided t-SNE. In the case of guided t-SNE, the coordinates of the resulting subtopics are consistent with the global view; this is not the case with standard t-SNE. In detail, guided t-SNE still places the subtopics from the orange topic at the top-right part within the lens, which is consistent with the coordinates of the original orange topic.

- DH-NMF: Dynamic hierarchical rank-2 NMF, which we proposed in Section 3.2.

Since the main idea of DH-NMF is to accelerate the process of building a topic hierarchy by utilizing initially built topic clusters, the number of initial topic clusters captured in the lens is a critical factor to DH-NMF. Therefore, we randomly selected four different document subsets with different numbers of initial topic clusters: 3, 5, 7, and 9. In all the experiments, we set the final number of topics, k_s, to generate inside the lens to 10. In addition, for each case, we obtained 300 and 600 documents from the NYT and VisPub datasets, respectively, to analyze how the number of documents affected the running time.

The comparisons of the computing times are shown in Fig. 4. Each result in this figure indicates the average value over 100 trials.

In all the cases, H-NMF and DH-NMF were shown to perform much faster than LDA. In addition, between H-NMF and DH-NMF, our proposed method, DH-NMF, performed better than the other. Although the performance gap between H-NMF and DH-NMF was relatively small when the number of initial topic clusters captured in the lens was small, e.g., 3 or 5, this gap grew with the increasing number of initial topic clusters. This result demonstrates the superiority of TopicLens based on DH-NMF in serving real-time topic modeling, and, moreover, as the number of initial topic clusters became larger, TopicLens updated the topic modeling results much more efficiently than the other methods.

4.2.2 Approximate t-SNE

To analyze the effectiveness of approximate t-SNE, we measured its computing times with respect to different landmark ratio values r. As discussed in Section 3.3, the value of r governs the total amount of computations in the main steps of t-SNE, which involves the computations of the cost function and the gradient vectors. Like in the previous experiment, we used the NYT and VisPub datasets and randomly selected 400 and 800 documents from each dataset, respectively.

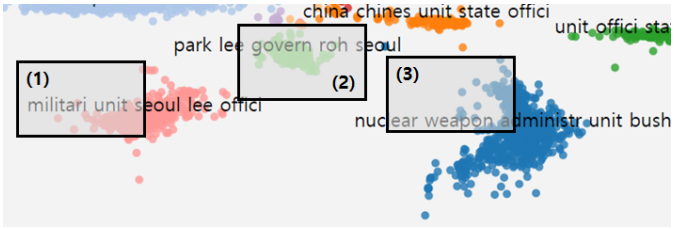
The results of the computing times from this experiment are shown in Fig. 5. Each value in this figure indicates the average value over 100 trials. Since approximate t-SNE with a landmark ratio equal to 1, i.e., $r = 1$, is equivalent to standard t-SNE, we report the computing times for the range of r from 0.2 to 0.8. On the other hand, the results computed by standard t-SNE, or, equivalently, approximate t-SNE with $r = 1$, are shown as a red line in Fig. 5. As can be seen in the figure, the computing time becomes smaller as the landmark ratio r gets bigger, and the amount of computing time saving is approximately proportional to $(1 - r)$.

However, a potential concern is that, as fewer data points are considered, the quality of 2D embedding might deteriorate. To resolve this issue, we conducted another experiment that analyzed the 2D embedding example of the t-SNE result, depending on a different landmark ratio r. For this experiment, we used the VisPub dataset. Fig. 6 shows that approximate t-SNE does not significantly impact the outcome of t-SNE. When $r = 0.2$, the topic clusters in the lens were less compact, but when $r = 0.6$, the result obtained was similar to that of the case when $r = 1$. Therefore, although the landmark ratio and the 2D embedding quality are in a trade-off relationship, with a suitable value of r, e.g., from 0.3 to 0.6, the user can achieve a reasonable quality of 2D embedding with much faster response time in TopicLens.

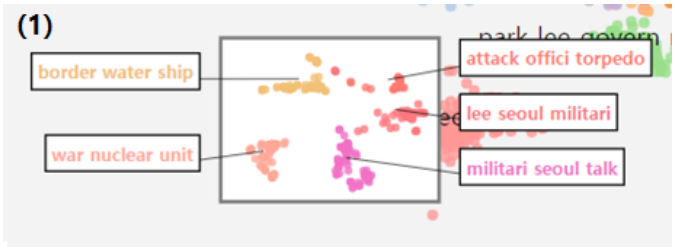
4.3 Consistency with the Global View

To validate the behavior of guided t-SNE in terms of consistency with the global view, we performed guided t-SNE and standard t-SNE on the same dataset (the VisPub dataset).

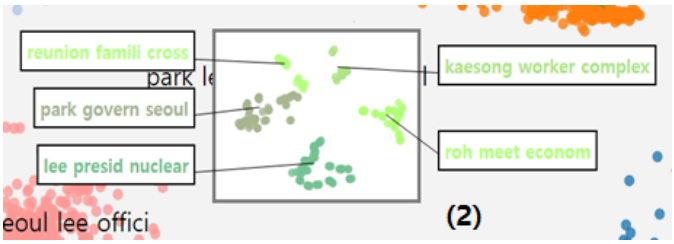
The comparison results are shown in Fig. 7. In this example, we applied TopicLens on the orange and blue topic clusters. Fig. 7(b) shows the visualization result of TopicLens when standard t-SNE was applied. As can be seen in this figure, the coordinates of the resulting sub-clusters are determined regardless of the initial coordinates of their parent clusters. For instance, the document subset from the orange-colored cluster, denoted as group (1), is placed across the two original topic clusters. Accordingly, the same phenomenon is found in group (2) as well. However, in Fig. 7(c), which shows the visualization of guided t-SNE, the coordinates of the newly computed sub-clusters



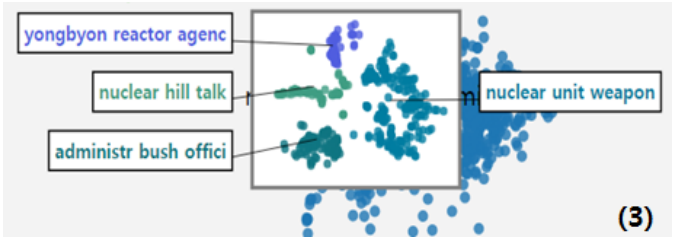
(a) Initial topic modeling.



(b) TopicLens result from area (1).



(c) TopicLens result from area (2).



(d) TopicLens result from area (3).

Fig. 8. Example topics revealed by TopicLens in the NYT dataset.

show more consistency with the global overview, allowing users to quickly recognize the subtopic structure produced by DH-NMF. For example, a sub-cluster, denoted as group (3), is placed near its original topic cluster; likewise, group (4) also exhibits a coherent placement with its original topic cluster. In terms of making sense of subtopic keywords, they are shown near the original corresponding topic cluster, and, thus, the result from guided t-SNE reduces the cognitive load of the users, which is caused by having to match the sub-clusters and their parent clusters. Therefore, this nice behavior of guided t-SNE allows TopicLens to effectively support a user's information needs by providing a 2D embedding consistent with the global view.

5 Usage Scenarios

Here, we present two usage scenarios demonstrating the dynamic topic modeling capability of TopicLens. In particular, we analyze the two datasets used in our experiments above.

5.1 New York Times Articles

We analyzed the NYT dataset collected from the search query "North Korea." Fig. 8(a) presents a part of an initial topic modeling visualization. In this visualization, we applied TopicLens to the three parts of the initial scatterplot, as shown in Fig. 8(a). First, we analyzed area (1), which revealed keywords such as "militari," "unit," and "seoul." Since we could not obtain a detailed information about a potential event based only on these initial keywords, we applied TopicLens to the part of the pink topic cluster, which revealed some salient topics and informed us that these documents are related to the incident involving a South Korean warship attacked by North Korea. For example, the words "attack" and "torpedo" gave clues about this incident. Additionally, other keywords, such as "lee," "militari,', and "talk," indicate the official announcement that President Lee made about this incident.

Second, area (2) containing the initial topics of "Park," "Lee," and "Roh," who are former presidents of South Korea, turned out to be a set of articles describing the relationship between South and North Korea. By exploring the topic keywords provided by TopicLens, we obtained information about a series of events such as "family reunion," which discussed the reunion of family members who were separated by the Korean War, and "economic cooperation" from the keywords "President Roh," "econom," and "Kaesong," where the last item corresponds to the Kaesong complex, a symbol of economic collaboration between South and North Korea. These events imply an amicable relationship between both countries. On the other hand, the keywords "President Lee" and "nuclear" indicate a "nuclear test" carried out by North Korea, representing a hostile relationship.

Lastly, area (3) revealed a set of keywords such as "nuclear," "weapon," and "Bush." On the basis of these keywords, we assumed that these articles are about the first nuclear test conducted by North Korea during the Bush administration. However, as shown in Fig. 8(d), newly discovered keywords by TopicLens such as "yongbyon" and "reactor" informed us that the nuclear test was related to the Yongbyon nuclear facility located in North Korea.

5.2 Academic Papers in the Areas of Visualization

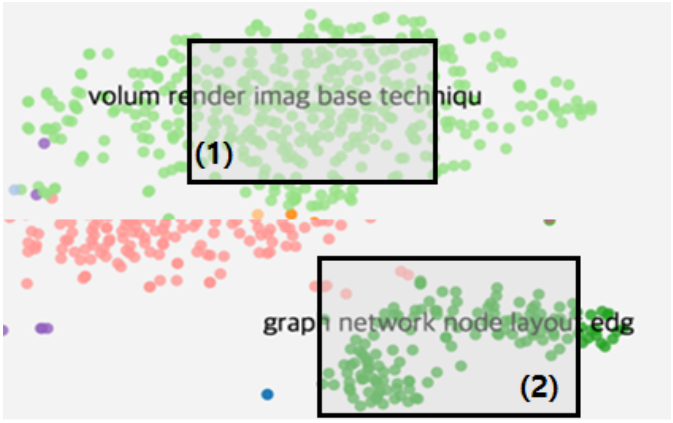
We also extracted meaningful topics from VisPub academic papers as follows. As shown in Fig. 9(a), we analyzed mainly two topic clusters revealing key topics such as "volume," "render," "graph," and "network." To further explore detailed information related to these research areas of visualization, we applied TopicLens to area (1). As shown in Fig. 9(b), while the progressive visualization was being performed, two subtopics from the intermediate output from DH-NMF were shown as "image rendering" and "volume rendering." Subsequently, DH-NMF further divided the cluster about "volume rendering" into two subtopics, with one containing "hardware" and the other containing "ray." After checking the detailed documents corresponding to these subtopics, we found that the subtopic of "volume, render, hardware" is mainly about the hardware acceleration in volume rendering, an active research topic in this area. On the other hand, the other sub-cluster containing "ray" turned out to be related to "volume ray casting," the technique that generates 2D images from 3D volumetric data.

Fig. 9(c) shows the TopicLens result applied to area (2), which contains the green-colored topics of "graph" and "network" and a small part of the pink-colored topic. When we applied TopicLens to this area, some meaningful keywords such as "social" and "tree" emerged, which corresponded to the research areas of treemap, tree layout, and social network. By examining the document details shown in the tooltip text, we found several articles on this subject, e.g., a research paper titled as "Using SocialAction to uncover structure in social networks over Time."

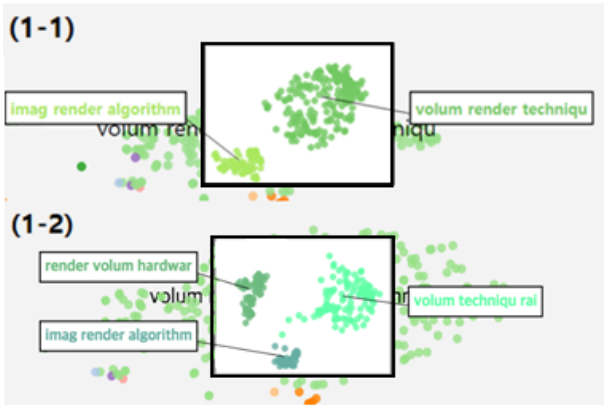
6 Discussion

The main novelty of our work lies in an effective integration of computational methods with a highly dynamic lens interface in the context of topic modeling. Such an integration can be further extended in the following aspects: (1) backend computational methods used in a main/initial view vs. those inside a lens and (2) frontend visualization methods used in an initial view vs. those inside a lens.

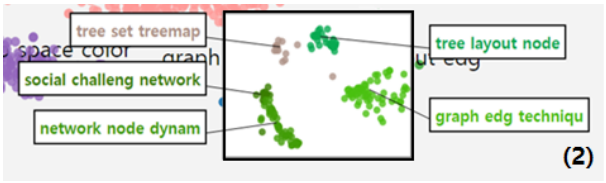
In this integration framework, TopicLens can be viewed as an example of using topic modeling in the backend and scatterplots in the frontend commonly in both an initial view and a lens. Using the same



(a) Initial topic modeling.



(b) TopicLens result from area (1).



(c) TopicLens result from area (2).

Fig. 9. Example topics revealed by TopicLens in the VisPub dataset.

computational and the same visualization methods, the user can maintain consistent perspectives about both an analytical and a visualization approaches. In other words, TopicLens allows the user to obtain subtopic information, which is the same type of information shown in a main view, but at a detailed level. On the other hand, using the same type of visualizations, e.g., scatterplots, inside and outside a lens, we can avoid any additional cognitive load on the user, which is caused when transitioning from understanding one type of visualization to another.

While not limited to topic modeling and scatterplots, it is possible to adopt different types of computational as well as visualization methods for various purposes. For example, an outlier detection method in the backend can be utilized inside a lens so that local outliers corresponding specifically to those data items inside a lens can be dynamically revealed. Similarly, instead of a scatterplot, different visualization types, such as treemaps or heatmaps, can be used to minimize the visual clutter due to the small screen space of a lens. Furthermore, stream-graph visualization can be adopted in either a main view or a lens to show the temporal trend of topics.

In all these extensions, one of the key requirements is the real-time support of computational methods against dynamically changing subsets of data. When a computational method requires intensive computational time, one potential solution would be to precompute the results on each of the possible data subsets. However, this is not always a perfect solution since we cannot prepare precomputed results for all the possible data subsets that a user may generate. For example, as seen in area (2) of Fig. 9(a) and its corresponding result shown in Fig. 9(c), the generated subtopics involve arbitrarily captured documents from each topic cluster, and, in this case, even if the full hierarchy of topics had been precomputed, its corresponding subtopics, which are generated based on the entire documents in each topic cluster, would not faithfully reflect such dynamically captured document subsets. Alternatively, similar to our proposed approach, the efficient on-demand computation by recycling the previously computed results can be an effective remedy to this issue. One may even think of a hybrid approach that combines the two complementary approaches of precomputation and on-demand computation. In this respect, our work can open up a wide range of possibilities in this research direction.

7 Conclusion and Future Work

We have presented a novel lens interface called TopicLens, which provides real-time topic modeling capabilities given a dynamically changing subset of documents captured in a lens. To this end, we proposed two new algorithms called dynamic hierarchical rank-2 nonnegative matrix factorization (DH-NMF) for topic modeling and guided approximate t-SNE for 2D embedding. TopicLens addresses two primary issues involved when integrating computational methods with visual analytics: significant computing time and non-interactivity, which prevent a user from obtaining fine-grained information in a visual analytic environment. As demonstrated in our quantitative results and usage scenarios, TopicLens helps a user interactively explore the user-specified subsets of data in real time, which delivers crucial knowledge that the initial run of topic modeling cannot provide.

Moreover, as discussed in Section 6, the idea of supporting highly dynamic interactions using computational methods can be further extended to other types of computational and visualization methods. Following this direction, we plan to build an advanced system that provides a diverse set of computational and visualization methods that users can choose within our dynamic lens interface. In addition, we plan to further improve the efficiency of computational methods by modifying the advanced methods such as Barnes-Hut t-SNE [45], which provides another efficient approximation of t-SNE.

Acknowledgments

Research reported in this publication was partially supported by NIH grant R01GM114267 and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2016R1C1B2015924). Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] C. Appert, O. Chapuis, and E. Pietriga. High-precision magnification lenses. In *Proc. the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 273–282, 2010.

[2] E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose. Toolglass and magic lenses: the see-through interface. In *Proc. the ACM Conference on Computer Graphics and Interactive Techniques*, pages 73–80, 1993.

[3] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.

[5] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu. FacetAtlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 16(6):1172–1181, 2010.

[6] M. S. T. Carpendale and C. Montagnese. A framework for unifying presentation space. In *Proc. the ACM Symposium on User Interface Software and Technology (UIST)*, pages 61–70, 2001.

[7] A. J.-B. Chaney and D. M. Blei. Visualizing topic models. In *Proc. the International Conference on Web and Social Media (ICWSM)*, pages 419–422, 2012.

[8] J. Choo, C. Lee, H. Kim, H. Lee, C. K. Reddy, B. L. Drake, and H. Park. PIVE: Per-iteration visualization environment for supporting real-time interactions with computational methods. In *Proc. the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 241–242, 2014.

[9] J. Choo, C. Lee, C. K. Reddy, and H. Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 19(12):1992–2001, 2013.

[10] J. Choo and H. Park. Customizing computational methods for visual analytics with big data. *IEEE Computer Graphics and Applications (CG&A)*, 33(4):22–28, 2013.

[11] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proc. the ACM Conference on Advanced Visual Interfaces (AVI)*, pages 74–77, 2012.

[12] J. Chuang, C. D. Manning, and J. Heer. "Without the clutter of unimportant words": Descriptive keyphrases for text visualization. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3):19, 2012.

[13] A. Cockburn, A. Karlson, and B. B. Bederson. A review of overview+detail, zooming, and focus+context interfaces. *ACM Computing Surveys*, 41(1):2, 2009.

[14] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. TextFlow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 17(12):2412–2421, 2011.

[15] V. De Silva and J. B. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford University, 2004.

[16] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41:391–407, 1990.

[17] M. Dumas, M. J. McGuffin, and P. Chasse. VectorLens: Angular selection of curves within 2D dense visualizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 21(3):402–412, 2015.

[18] N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Color Lens: Adaptive color scale optimization for visual exploration. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 17(6):795–807, 2011.

[19] Fekete, Jean-Daniel and Plaisant, Catherine. Excentric labeling: Dynamic neighborhood labeling for data visualization. In *Proc. the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 512–519, 1999.

[20] D. Fisher, I. Popov, S. Drucker, and M. C. Schraefel. Trust me, I'm partially right: Incremental visualization lets analysts explore large datasets faster. In *Proc. the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1673–1682, 2012.

[21] G. W. Furnas. Generalized fisheye views. In *Proc. the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 16–23, 1986.

[22] G. H. Golub and C. F. van Loan. *Matrix Computations, third edition*. Johns Hopkins University Press, Baltimore, 1996.

[23] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth. TopicNets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):23, 2012.

[24] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 8(1):9–20, 2002.

[25] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 50–57, 1999.

[26] T. Iwata, T. Yamada, and N. Ueda. Probabilistic latent semantic visualization: topic model for visualizing documents. In *Proc. the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 363–371, 2008.

[27] W. Javed, S. Ghani, and N. Elmqvist. PolyZoom: multiscale and multifocus exploration in 2D visual spaces. In *Proc. the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 287–296, 2012.

[28] J. Kim, Y. He, and H. Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

[29] J. Kim and H. Park. Sparse nonnegative matrix factorization for clustering. 2008.

[30] D. Kuang and H. Park. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In *Proc. the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 739–747, 2013.

[31] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[32] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. iVisClustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum (CGF)*, 31(3pt3):1155–1164, 2012.

[33] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 497–506, 2009.

[34] Y. K. Leung and M. D. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 1(2):126–160, 1994.

[35] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo. TopicPanorama: a full picture of relevant topics. In *Proc. theof the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 183–192, 2014.

[36] E. Pietriga and C. Appert. Sigma lenses: focus-context transitions combining space, time and translucence. In *Proc. the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1343–1352, 2008.

[37] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proc. the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 569–577, 2008.

[38] R. Rao and S. K. Card. The Table Lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In *Proc. the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 318–322, 1994.

[39] G. G. Robertson and J. D. Mackinlay. The Document Lens. In *Proc. the ACM Symposium on User Interface Software and Technology (UIST)*, pages 101–108, 1993.

[40] M. Sarkar and M. H. Brown. Graphical fisheye views of graphs. In *Proc. the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 83–91, 1992.

[41] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. the IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[42] C. D. Stolper, A. Perer, and D. Gotz. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(12):1653–1662, 2014.

[43] C. Tominski, J. Abello, F. Van Ham, and H. Schumann. Fisheye tree views and lenses for graph visualization. In *Proc. the International Conference on Information Visualization (InfoVis)*, pages 17–24, 2006.

[44] C. Tominski, S. Gladisch, U. Kister, R. Dachselt, and H. Schumann. A survey on interactive lenses in visualization. In *State of the Art Reports for the European Conference on Visualization*, 2014.

[45] L. Van Der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of machine learning research (JMLR)*, 15(1):3221–3245, 2014.

[46] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(2579–2605):85, 2008.

[47] F. van Ham and J. J. van Wijk. Interactive visualization of small world graphs. In *Proc. the IEEE Symposium on Information Visualization (InfoVis)*, pages 199–206, 2004.

[48] C. Ware and M. Lewis. The DragMag image magnifier. In *Conference Companion of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 407–408, 1995.

[49] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. TIARA: a visual exploratory text analytic system. In *Proc. the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 153–162, 2010.

[50] N. Wong, S. Carpendale, and S. Greenberg. EdgeLens: An interactive method for managing edge congestion in graphs. In *Proc. the IEEE Symposium on Information Visualization (InfoVis)*, pages 51–58, 2003.

[51] P. C. Wong, H. Foote, D. Adams, W. Cowley, and J. Thomas. Dynamic visualization of transient data streams. In *Proc. the IEEE Symposium on Information Visualization (InfoVis)*, pages 97–104, 2003.