# UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization

Jaegul Choo, Changhyun Lee, Chandan K. Reddy, and Haesun Park

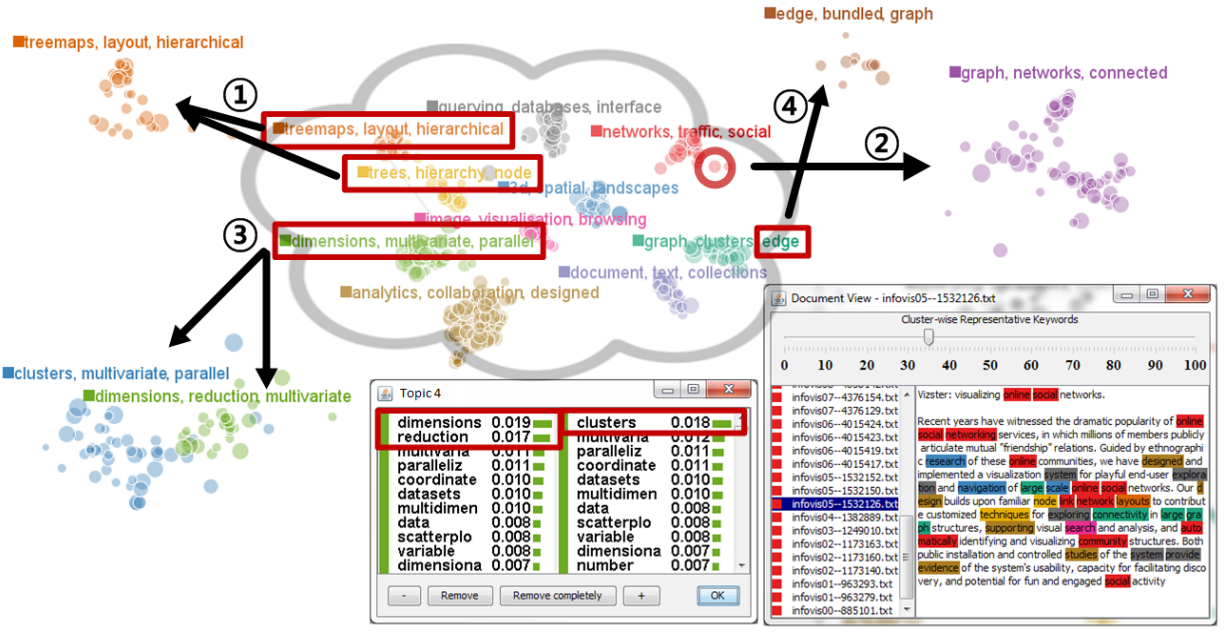



Fig. 1. An overview of UTOPIAN. Given a scatter plot visualization generated by the modified t-SNE, UTOPIAN provides various interaction capabilities: (1) topic merging, (2) document-induced topic creation, (3) topic splitting, and (4) keyword-induced topic creation. Additionally, the user can refine topic keyword weights. The document viewer highlights the representative keywords from each topic.

**Abstract**—Topic modeling has been widely used for analyzing text document collections. Recently, there have been significant advancements in various topic modeling techniques, particularly in the form of probabilistic graphical modeling. State-of-the-art techniques such as Latent Dirichlet Allocation (LDA) have been successfully applied in visual text analytics. However, most of the widely-used methods based on probabilistic modeling have drawbacks in terms of consistency from multiple runs and empirical convergence. Furthermore, due to the complicatedness in the formulation and the algorithm, LDA cannot easily incorporate various types of user feedback. To tackle this problem, we propose a reliable and flexible visual analytics system for topic modeling called UTOPIAN (**U**ser-driven **Top**ic modeling based on **I**ntera**c**tive **N**onnegative Matrix Factorization). Centered around its semi-supervised formulation, UTOPIAN enables users to interact with the topic modeling method and steer the result in a user-driven manner. We demonstrate the capability of UTOPIAN via several usage scenarios with real-world document corpuses such as InfoVis/VAST paper data set and product review data sets.

**Index Terms**—Latent Dirichlet allocation, nonnegative matrix factorization, topic modeling, visual analytics, interactive clustering, text analytics

## 1 Introduction

Due to an ever increasing amount of document data and complexities involved with analyzing them in practice, revealing meaningful insights and thus guiding users in their decision-making processes has long been an active area of research. Among various approaches to tackle these problems, a topic modeling approach, which discovers semantically meaningful topics from a document corpus, has been gaining popularity in both the fields of data mining/machine learning and visual analytics.

Similar to many other techniques dealing with document data, most of these topic modeling approaches take an input in the form of a term-document matrix representation of documents via a bag-of-words model. Different kinds of definitions about topic modeling may exist, but from a practical standpoint, the topic modeling approaches typically give two types of outputs: (1) a representation of each topic


- *Jaegul Choo is with Georgia Institute of Technology. E-mail: jaegul.choo@cc.gatech.edu.*
- *Changhyun Lee is with Georgia Institute of Technology. E-mail: clee407@gatech.edu.*
- *Chandan K. Reddy is with Wayne State University. E-mail: reddy@cs.wayne.edu.*
- *Haesun Park is with Georgia Institute of Technology. E-mail: hpark@cc.gatech.edu.*


*Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.*
*For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.*

in terms of a weighted combination of keywords, or a *keyword-wise topic representation* in short, and (2) a representation of each document in terms of a weighted combination of topics, or a *topic-wise document representation* in short. In the two representations, the weight indicates how closely a particular keyword is related to the corresponding topic/document and how closely a particular topic is related to the corresponding document, respectively.

From this perspective, topic modeling is related to soft clustering where the documents are represented as weighted combinations of clusters in terms of their proximity to each cluster, which essentially play the same role of the second output of topic modeling. However, soft clustering typically tends to focus only on how closely a particular document is related to each cluster while topic modeling deals with both types of outputs including the semantic meaning of each cluster/topic on its own. The work presented in this paper deals with the latter case, and we primarily present our system in the context of topic modeling.

Allowing some relaxation on topic modeling definitions by involving possibly negative weights on the topic modeling output, the history of topic modeling approaches traces back to a well-known traditional technique called latent semantic indexing (LSI) [11]. However, researchers and end-users have had difficulties in making sense out of negative weights on keywords or topics, which prevent LSI from being used in real-world domains. Alternatively, more recent methods have focused on probabilistic models that have nice interpretation properties in their outputs since both of the above-mentioned outcomes are all non-negative and are summed up to one under a probabilistic framework. These probabilistic topic modeling methods, such as probabilistic latent semantic indexing (p-LSI) [17] and latent Dirichlet allocation (LDA) [4], have become popular in various domains. Especially, LDA has also been widely applied in visual analytics domain primarily due to its excellent performances compared to most of the previous approaches.

However, when applied in visual analytics, LDA has several practical shortcomings in terms of *consistency from multiple runs* and *empirical convergence*. As will be described in Section 6, the former indicates how stable the algorithm output remains from multiple runs with the same setting while the latter indicates how early the algorithm converges from a user's point of view compared to algorithmic convergence. Furthermore, due to the complicatedness in the formulation and the algorithm, incorporating various types of user feedback with LDA is relatively difficult.

In order to overcome these drawbacks, we propose a reliable and flexible topic modeling visual analytics system called UTOPIAN (**U**ser-driven **Top**ic modeling based on **I**nter**a**ctive **N**onnegative Matrix Factorization). Nonnegative matrix factorization (NMF) [30, 25] has been one of the most active research areas in data mining and machine learning fields, and it has been applied in the context of topic modeling mainly from a computational perspective [3]. As an approach for topic modeling, NMF works similar to LSI in that they both solve a matrix decomposition problem given a particular rank value corresponding to the number of topics. However, as the name suggests, NMF imposes non-negativity constraints on every element of the resulting matrices so that it can maintain interpretability.

The advantage of UTOPIAN is that NMF does not suffer from the previously raised issues while providing results with comparable quality to those of LDA. In other words, although NMF is non-convex similar to LDA, its algorithm usually generates *a consistent result from multiple runs* for a given document corpus (Section 6). Moreover, the NMF algorithm is *deterministic*. Thus, unless the user modifies an initial specification, she will obtain the same result from the algorithm. These desirable behaviors of NMF serve as important grounds to make UTOPIAN practically useful and interactive in real-world visual analytics by enabling the user to progressively improve a particular result by interactively changing the algorithm specifications, etc.

More importantly, another notable advantage of UTOPIAN is that it can easily incorporate more active user interactions that are beyond changing parameters, initial values, etc., via various forms of semi-supervisions on NMF. Thus, UTOPIAN provides these flexible interaction capabilities of NMF in improving the topic modeling result in a user-driven manner. The manner in which the adopted semi-supervised NMF method takes the user interventions into account is intuitive because the semi-supervision will be in the same form as the two above-described topic modeling outputs which the user is already familiar with throughout his/her analysis. This characteristic removes any additional need for transforming the user interventions back to the algorithm parameters or constraints in an ambiguous way.

Based on the semi-supervised NMF method, UTOPIAN provides a wide variety of interaction capabilities for improving topic modeling. UTOPIAN visualizes the NMF topic modeling result mainly in a node-link diagram by using a variant of one of the state-of-the-art dimension reduction methods called t-distributed stochastic neighborhood embedding (t-SNE) [34]. The provided interaction forms primarily two types of semi-supervisions to the algorithm: keyword-based and document-based ones, each of which can also be flexibly generated in either a top-down or a bottom-up manner (Section 4). Furthermore, in order to support real-time interactions with the topic modeling algorithm, UTOPIAN visualizes its intermediate outputs even before its convergence, (Section 5.2). Finally, by using UTOPIAN, we show several interesting usage scenarios where the topic modeling results are interactively improved.

The main contributions of UTOPIAN are summarized as follows:

- Proposing NMF as a better alternative topic modeling method in visual analytics compared to LDA.
- Developing a visual analytics system called UTOPIAN equipped with various user interaction capabilities based on the semi-supervised NMF.
- Presenting various usage scenarios using real-world data in which the topic modeling result is improved via various user interactions.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 introduces NMF in the context of topic modeling. Section 4 presents the user interactions for improving topic modeling using NMF. Section 5 describes UTOPIAN supporting these user interactions. Section 6 shows detailed quantitative results about the algorithm behaviors of LDA and NMF. Section 7 presents several usage scenarios with UTOPIAN. Finally, Section 8 concludes the paper with some future work.

## 2 Related Work

The primary goal of discovering topics in document collections is to provide the user with a summary of a document corpus in terms of, for example, keyword summaries of computed topics and a group of documents closely related to each topic. Hence, topic modeling can be viewed as a (soft) clustering in the sense that clustering also gives groups of semantically coherent documents where the semantic meaning of each cluster can be represented as the most frequent keywords in the document group. Therefore, our discussion in this section encompasses previous studies involving topic modeling as well as document clustering in visual analytics.

A well-known visual document analysis system, IN-SPIRE [36], shows the summaries of the topic clusters using standard clustering and dimensionality algorithms such as $k$-means and principal component analysis (PCA) [21], respectively, to compute topic summaries. However, IN-SPIRE does not support substantial interactions for improving topic clusters. Another widely-used system, Jigsaw [33], provides various capabilities such as summarizing document cluster views and allows several basic interactions such as changing the number of clusters and initializing seed documents of clusters. More recently, various visual analytics systems involving an advanced topic modeling method, LDA, have been proposed. TIARA [35], one of the first systems that applied LDA in visual analytics, has utilized a ThemeRiver-style visualization to show the temporal trend of topic evolution. Other work such as ParallelTopics [12] and TextFlow [10] have also aimed

Table 1. Notations in the Paper

| Notation | Description |
|---|---|
| $m$ | the number of keywords |
| $n$ | the number of documents |
| $k$ | the number of topics |
| $X \in \mathbb{R}_+^{m\times n}$ | A term-by-document matrix |
| $W \in \mathbb{R}_+^{m\times k}$ | A term-by-topic matrix |
| $H \in \mathbb{R}_+^{k\times n}$ | A topic-by-document matrix |
| $w_l \in \mathbb{R}_+^{m\times 1}$ | A keyword-wise representation of the $l$-th topic |
| $h_i \in \mathbb{R}_+^{k\times 1}$ | A topic-wise representation of document $i$ |
| $V \in \mathbb{R}_+^{m\times k}$ | A reference term-by-topic matrix for $W$ |
| $G \in \mathbb{R}_+^{k\times n}$ | A reference topic-by-document matrix for $G$ |
| $v_l \in \mathbb{R}_+^{m\times 1}$ | A reference vector for $w_l$ |
| $g_i \in \mathbb{R}_+^{k\times 1}$ | A reference vector for $h_i$ |
| $M_W \in \mathbb{R}_+^{k\times k}$ | A mask/weight matrix for the columns of $W$ |
| $M_H \in \mathbb{R}_+^{n\times n}$ | A mask/weight matrix for the columns of $H$ |
| $M_W^{(l)}$ | A mask/weight value for $w_l$ |
| $M_H^{(i)}$ | A mask/weight value for $h_i$ |

at capturing the topic changes and understanding the document characteristics based on LDA. Specifically, TextFlow handles merging and splitting topics over time as well as detects emerging/diminishing topics. In most of these studies, however, the main focus lies in how effectively the LDA results can be utilized in their applications, but not the interactions with a topic modeling itself for improving its result. As noted in recent work in the machine learning domain [1, 20], incorporating domain knowledge into topic modeling via user interactions can be a cumbersome process.

On the other hand, there have been many studies that have tried to improve clustering or topic modeling results in document analysis through user interactions. Such interactive clustering approaches go well beyond the standard document analysis [18, 29, 5]. The importance and the need for an interactive visual exploration for text document collection have also been gaining a lot of interest in recent years [13, 16, 15]. Similar to our approach, several studies have actively used node-link diagrams, which visualize documents along with their clusters, allowing users to interactively create a hierarchical structure of topic clusters [27, 31]. iCluster [14] lets the user manually perform clustering from scratch based on the recommended documents computed by the system. Although it may be inefficient in large-scale data, the user can maintain semantically meaningful topic clusters, and the system can reflect user feedback at every step of performing clustering for individual documents. iVisClustering [26] is another visual analytics system for document clustering using LDA. It provides a few capabilities to directly interact with LDA, such as manipulating the topical keyword weights and merging/splitting topic clusters, etc. However, most of these LDA-based systems suffer from the previously discussed problems, and due to the significant running time of LDA, most computations are done off-line, and real-time interactions with LDA is difficult. In contrast, NMF, which does not suffer from these problems, has never been systematically explored in the context of visual analytics. In this regard, UTOPIAN is one of the first such systems based on NMF for document topic modeling that supports various user interactions in real-time visual environments.

## 3 Nonnegative Matrix Factorization (NMF) for Topic Modeling

In this section, we introduce NMF in the context of topic modeling and compare it with widely-used probabilistic topic modeling. The notations used in the paper is summarized in Table 1.

### 3.1 Formulation

Given a nonnegative matrix $X \in \mathbb{R}_+^{m\times n}$, and an integer $k \ll \min(m,n)$, NMF finds a lower-rank approximation given by

$$X \approx WH, \tag{1}$$

where $W \in \mathbb{R}_+^{m\times k}$ and $H \in \mathbb{R}_+^{k\times n}$ are nonnegative factors. NMF is typically formulated in terms of the Frobenius norm as

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

where $' \geq '$ applies to every element of the given matrix in the left-hand side. In the topic modeling context, $x_i \in \mathbb{R}_+^{m\times 1}$, the $i$-th column of $X$, corresponds to the bag-of-words representation of document $i$ with respect to $m$ keywords, possibly with some pre-processing, e.g., inverse-document frequency weighting and column-wise $l_2$-norm normalization. $k$ corresponds to the number of topics. $w_l \in \mathbb{R}_+^{m\times 1}$, the $l$-th nonnegative column vector of $W$, represents the $l$-th topic as a weighted combination of $m$ keywords. A large value indicates a close relationship of the topic to the corresponding keyword. $h_i \in \mathbb{R}_+^{k\times 1}$, the $i$-th column vector of $H$, represents document $i$ as a weighted combination of $k$ topics.

### 3.2 NMF for Topic Modeling

Compared to standard topic modeling methods such as p-LSI and LDA, NMF essentially gives the same output types, (1) a keyword-wise topic representation, e.g., $w_l$, and (2) a topic-wise document representation, e.g., $h_i$. The only difference, however, is that $w_l$ and $h_i$ are not necessarily column-normalized (meaning summing up to one) unlike the p-LSI and LDA outputs. Nonetheless, such a difference is negligible in that Eq. (1) can be manipulated via diagonal scaling matrices as

$$X \approx WH = (WD_W)\left(D_W^{-1}H\right) = \hat{W}\hat{H}$$

where the $(l,l)$-th entry of the diagonal matrix $D_w \in \mathbb{R}_+^{k\times k}$ corresponds to the sum of $w_l$. Now the new matrix $\hat{W}$ is column-normalized, giving an equivalent output to the first outputs from p-LSI and LDA, but the second output $\hat{H}$ is still not column-normalized. The column normalization on $\hat{H}$ does not affect the interpretation of each document in terms of its relative relationships to topics. In this sense, NMF can be used as an alternative to standard topic modeling methods.

## 4 User Interactions via Semi-supervised NMF

In this section, we first describe a semi-supervised NMF (SS-NMF) that can flexibly support various user interactions in UTOPIAN. We then propose a set of user interactions and describe how they are performed via SS-NMF formulations.

### 4.1 Semi-supervised NMF (SS-NMF)

In addition to the original NMF inputs, $X$ and $k$, SS-NMF takes additional inputs as reference matrices, $V \in \mathbb{R}_+^{m\times k}$ and $G \in \mathbb{R}_+^{k\times n}$ for $W$ and $H$, respectively, and diagonal matrices $M_W \in \mathbb{R}_+^{k\times k}$ and $M_H \in \mathbb{R}_+^{n\times n}$, which assign weights on the columns of $V$ and $G$, respectively. Given these inputs, SS-NMF includes additional terms that penalize the difference between $G$ and $H$ (up to row-wise scaling via $D_H$) and that between $V$ and $W$ as

$$\begin{aligned}\min_{W,H,D_H\geq 0} \quad & \Big\{ \|X - WH\|_F^2 \\ & + \|(W - V)M_W\|_F^2 + \|(H - GD_H)M_H\|_F^2 \Big\}\end{aligned} \tag{3}$$

for nonnegative factors $W \in \mathbb{R}_+^{m\times k}$ and $H \in \mathbb{R}_+^{k\times n}$ and a diagonal matrix $D_H \in \mathbb{R}_+^{n\times n}$.

Basically, Eq. (3) regularizes/supervises the resulting $W$ and $H$ to be as close as possible to $V$ and $G$, respectively, while still approximating the input matrix $X$ as $WH$. More specifically, $w_l$ and $h_i$ are enforced to be close to $v_l \in \mathbb{R}^{m\times 1}$, the $l$-th column vector of $V$, and $g_i \in \mathbb{R}^{k\times 1}$, the $i$-th column vector of $G$, respectively. The diagonal matrix $D_H$ plays a role of automatically adjusting the scales between $h_i$'s and $g_i$'s. Note that as discussed in Section 3, $D_H$ does not change the relative topical weight values, and thus the effect of $D_H$ can be ignored when interpreting the topic modeling results of SS-NMF.

$M_W$ and $M_H$ enable such supervision to be applied selectively on a subset of columns of $W$ and $H$ when the corresponding diagonal entries of $M_W$ and $M_H$ are set to zeros. In other words, when $M_W^{(l)}$, the $(l, l)$-th entry of $M_W$, is set to zero, no supervision on $w_l$ is imposed. Likewise, when $M_H^{(i)}$, the $(i, i)$-th entry of $M_H$, is set to zero, no supervision on $h_i$ is imposed. On the other hand, larger diagonal values of $M_W$ and $M_H$ supervise more strongly the corresponding columns of $W$ and $H$, respectively.

When used in interactive topic modeling, the reference information represented in $V$ and $G$ represents the prior knowledge the user wants to impose in the topic modeling output. On the one hand, by setting only a few nonzero entries in $M_W$ and $M_H$ for partial supervision, the user can selectively regularize particular topics instead of the entire topics. On the other hand, by setting relatively small nonzero values in $M_W$ and $M_H$, the user can weakly supervise his/her knowledge into the formulation in case she is not completly sure about what a topic should look like.

### 4.2 Supported User Interactions

Even though the SS-NMF formulation provides a natural way to impose the user's prior knowledge in topic modeling processes, it is nontrivial to *design a semantically meaningful set of user interactions for improving topic modeling*. In this section, we propose a variety of such user interactions based on SS-NMF as follows: (i) keyword refinement of an existing topic, (ii) topic splitting/merging for interactive adjustment of the number of topics, and (iii) keyword-induced/document-induced topic creation. The proposed user interactions are mostly on an individual topic or document basis by generating an appropriate $v_l$ or $g_i$, respectively, with its corresponding nonzero weight values for $M_W^{(l)}$ and $M_H^{(i)}$. In the following, we describe each of the supported user interactions.

**Topic keyword refinement.** This interaction enables the user to change keyword weights so that she can directly refine the semantic meaning of a topic. That is, starting from $w_l$, the user can increase/decrease or even remove the weights of particular keywords and set the modified vector to $v_l$ along with a nonzero $M_W^{(l)}$. For instance, suppose $w_l$ represents the distribution over three keywords, 'apple,' 'orange,' and 'banana,' as (1.3, 0.6, 0.1). If the user wants to completely remove the term 'banana' from the meaning of this topic, she can set $v_l$ to (1.3, 0.6, 0). On the other hand, if she wants this topic to be more closely related to 'orange,' then she may set $v_l$ to (1.3, 1.6, 0.1). Such reference information affects the subsequent running of SS-NMF.

**Topic merging.** This interaction allows the user to merge similar topics into a single one. Our approach to achieve this interaction is to generate $g_i$ from $h_i$ as follows. We first identify the documents related most closely to either of the merged topics, which are essentially the documents that are hard-clustered to the merged topics. For these documents, we obtain their $h_i$'s and merge the values corresponding to the two merged topics by adding them up to a single value, and set $g_i$'s to the resulting $h_i$'s. For instance, suppose two documents, whose $h_i$'s are represented as (0.7, 0.2, 0.1) and (0.3, 0.5, 0.2), respectively, in terms of the three topics. When merging topic 1 and 2, the corresponding $g_i$'s would be set to (0.7+0.2, 0.1) and (0.3+0.5, 0.2), respectively, in terms of the merged topic and topic 3.

Notice that even though our merging algorithm is document-based, the keywords associated with the representation of the merged topics in terms of keywords are adjusted accordingly, which could potentially make the merged topic bring new documents in and/or exclude existing documents out.

**Topic splitting.** In this interaction, we provide a capability to split a topic into two in a user-driven way. In this interaction, we assume the user expect the two split topics share a common semantic meaning at a high level but with minor differences in their details. More specifically, suppose the user wants to split a particular topic $w_l$, into two topics $w_a$ and $w_b$. We first initialize the reference vectors $v_a$ and $v_b$ for the two split topics as $w_l$. Afterwards, we let users manipulate these two topic vectors via the *topic keyword refinement* interaction so that the splitting process can reflect the user's intention. For instance, given the same keyword-wise representations of two split topics, the user might want to increase the weight of a particular keyword in the first one while decreasing/removing the weight of the same keyword in the second one. Based on the reference information $v_a$ and $v_b$ generated in this manner, the subsequent running of SS-NMF performs the topic splitting process.

**Document-induced topic creation.** This interaction constructs a topic based on a small number of exemplar documents of the user's choice. In this interaction, for those documents specified as exemplars by the user, the corresponding $g_i$'s are initialized to zero vectors but are set to one for the value corresponding to the newly created topic. For instance, when the current number of topics is $k$, $g_i$'s are set to a $(k+1)$-dimensional vector where the first $k$ entries are zeros but the last entry is set to one, assuming the $(k+1)$-th topic is the newly created one. Such a process enforces the exemplar documents to be related purely to the newly created topic. Using this reference information, SS-NMF forms a new keyword-wise topic representation for a new topic, and accordingly the relevant documents to the exemplars.

**Keyword-induced topic creation.** This interaction provides a way to create a topic based on a small set of keywords of the user's interest. In this interaction, we assume that the user is given the topic summaries in terms of the keywords with the largest weight in the keyword-wise topic representation $w_l$. With this interaction, the user can select several interesting keywords and create a topic based on them. For its formulation using SS-NMF, we initialize the reference information $v_l$ of a newly created topic as a zero vector. Then, for the user-selected keywords, we set their corresponding values as ones in $v_l$. In this manner, the resulting $w_l$ is enforced to be related mainly to these keywords. As a result, the documents closely related to these keywords are included in this topic.

### 4.3 Perspectives of Supported User Interactions

So far, we have mainly presented various user interactions in terms of how they can be formulated as the reference information about the topic modeling outputs in SS-NMF. Depending on how we generate such reference information, we discuss the above-described user interactions from the following perspectives.

**Keyword-based vs. document-based.** The user interactions can be formulated using the reference information for either the keyword-wise topic representation or the topic-wise document representation. *Keyword-based interactions* refer to those manipulating the reference information $v_l$ about the keyword-wise topic representation $w_l$ of a particular topic. On the other hand, *document-based interactions* refer to those manipulating the reference information $g_i$ about the topic-wise document representation $h_i$ of a particular document. From this perspective, *topic keyword refinement*, *topic splitting*, and *keyword-induced topic creation* are keyword-based while *topic merging* and *document-induced topic creation* are document-based.

Nonetheless, in most topic modeling methods including NMF, the two main outputs of topic modeling are interdependent. In other words, a change of any one of them affects the other. Therefore, some interactions can be formulated in the other way. For example, *topic splitting* can be performed based on the user interaction of splitting the documents in one topic into two groups, which would then be a document-based formulation. In addition, topic merging may also be formulated as an averaged keyword-wise topic representation of the two, which would be keyword-based. However, after experimenting all these various options, we found the proposed approaches for the above interactions reflect the user intention properly in terms of the SS-NMF algorithm behavior.

**Template-based (top-down) vs. from-scratch-based (bottom-up).** The user interactions can also be characterized in terms of how we form the reference information about the topic modeling output. *Template-based interactions* refer to those starting from a current topic modeling output $w_l$ or $h_i$ (as a template) and manipulating it to generate the reference information $v_l$ or $g_i$. On the other hand, *from-scratch-based interactions* refer to those starting from a completely zero vector of the reference information $v_l$ or $g_i$ and putting nonzero values only in

the entries that the user specifies. From this perspective, *topic keyword refinement*, *topic merging*, and *topic splitting* are template-based while *document-induced topic creation* and *keyword-induced topic creation* are document-based.

In general, these two methods contrast in terms of the quality and the efficiency when generating the reference information via a user interaction. That is, the template-based approach has an advantage of being able to efficiently create the reference information by using the already-built topic modeling output, which would contain most keywords and topics with nonzero weight values in $w_l$ and $h_i$, respectively. After doing so, the user can gradually refine the reference information. This approach, however, may suffer from the poor quality of an initial topic modeling output.

On the other hand, the from-scratch-based method does not have this problem since the reference information do not involve any information from the topic modeling output. Instead, the reference information starts from a completely zero vector, and it is created entirely as what the user specifies. For instance, from the perspective of the SS-NMF algorithm, *topic keyword refinement* and *keyword-induced topic creation* has no difference in that both of them create the reference information $v_l$ about $w_l$. However, the latter, which is a from-scratch-based method, does not allow any keywords to be involved other than the user-specified keywords because we set nonzero values in $v_l$ only for the corresponding keywords. In this sense, the from-scratch-based method can maintain the quality of the reference information based on the user intention, but it may be inefficient to involve a large number of keywords or documents in the reference information since the user has to manually go through all the processes.

## 5 UTOPIAN

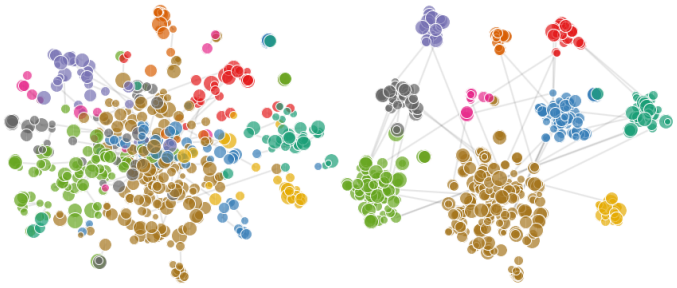
UTOPIAN[1] visually presents the SS-NMF result in a node-link diagram in which the displayed points represent individual documents with color-coding their (hard-clustered) topic cluster labels. Additionally, we provide the summary of topic clusters in terms of their most representative keywords. Given a visualization created in this manner, the user can perform the above-described interactions to improve the topic modeling result. UTOPIAN has three key features:

- Modified t-distributed stochastic neighborhood embedding (t-SNE) as a layout algorithm for proper visualization of documents and their topics
- Visualization of the intermediate algorithm outputs for real-time visualization and responsive interaction with NMF
- Animated visualization of explicit topic cluster changes for tracking the progression of the topic modeling outputs

### 5.1 Modified t-distributed Stochastic Neighborhood Embedding (t-SNE)

In a visual analytics approach, it is important for the user to be able to visually understand and interact with the topic modeling results. As our main visual layout algorithm to visualize documents along with their topic modeling outputs in a node-link diagram form, we have chosen one of the state-of-the-art methods called t-distributed stochastic neighborhood embedding (t-SNE) [34]. Given high-dimensional vectors of data items or their pairwise similarity values, t-SNE computes a 2D layout that reflects the high-dimensional relationships or the given pairwise similarities in terms of their 2D Euclidean distances. When applied in document data, their bag-of-words representations or their cosine similarity measures are typically used as an input to t-SNE.

The main reason we have chosen t-SNE rather than other standard techniques such as principal component analysis [21] and multidimensional scaling [23] is that t-SNE has shown its outstanding capabilities in revealing the implicit groupings of data items in visualization applications [34]. However, we found that t-SNE often generate a node-link diagram for document data where different topic clusters severely overlap, as shown in Fig. 2(a), making it difficult to analyze the topic-vs-topic and/or topic-vs-document relationships. To overcome this problem, we adopted a supervision idea in dimension reduction, such as linear discriminant analysis [19], which tries to represent the clear cluster structure in dimension reduction results. Although this kind of behavior may distort the original relationships of data items, it has been shown to be useful in visualization applications in various domains [6, 7].



(a) The original t-SNE (b) The modified t-SNE

Fig. 2. A comparison between the original and the modified t-SNE. The modified t-SNE (b) with a shrinkage parameter value of 0.4 shows a much clearer structure of topic clusters than the original t-SNE (a). UTOPIAN provides a slider interface to control the shrinkage parameter value. In addition, the edges have been drawn for the pairs of data points whose distances are closer than a user-specified threshold. 515 documents of the InfoVis-VAST data set have been used.

In detail, we have modified the original t-SNE algorithm so that they can better show the cluster structure in noisy data sets such as documents. Suppose $d_{ij}$ denotes the pairwise distance between documents $i$ and $j$ to be used as an input to t-SNE. Given a topic cluster index obtained by applying hard-clustering to the topic modeling result, if these two documents belong to the same topic cluster, then we decrease their pairwise distance by a factor of $\alpha$, i.e., $\alpha d_{ij}$ where $\alpha$ is a pre-defined parameter value between zero and one. By default, the shrinking parameter $\alpha$ is set to 0.4, but the user can interactively change it via a slider interface in UTOPIAN. In this manner, the pairwise distance input to t-SNE now represents each cluster more compactly, resulting in a clearer visualization of the cluster structure. Fig. 2(b) shows an example visualization when applying the proposed modification to t-SNE. As we can see, the topic clusters are shown much better than Fig. 2(a).

### 5.2 Visualization of the Intermediate Algorithm Outputs for Real-time Interactive Visualization

The real-time interaction capability is crucial in making continuous interactions efficiently based on highly exploratory nature in visual analytics. To achieve this goal, UTOPIAN is designed to reflect the intermediate algorithm outputs of NMF and t-SNE as soon as they are available [9]. Furthermore, once the user performs a particular interaction, the algorithm responds immediately to the user interaction from the immediate next iterations. In this manner, she can obtain its effects of the algorithm in real time. However, one potential drawback of this approach is that the visualization/rendering process may be burdensome since it has to be performed repetitively over the iterations instead of only once after the final iteration. To overcome this issue, we adopt a multi-threading approach to separate computational and visualization processes into different threads that can be efficiently executed simultaneously.

However, it is not straightforward to adopt this idea because UTOPIAN involves two computational modules and one of them (t-SNE) is dependent on the other (SS-NMF) due to the modification described in Section 5.1. Therefore, we further develop this idea so that it can handle two dependent computational modules simultaneously. Basically, we now have two separate threads for SS-NMF and t-SNE and another for visualization/rendering. Between the t-SNE and the visualization threads, the iteration-wise computational visualization

[1] http://fodava.gatech.edu/UTOPIAN

Table 2. The Summary of the Data Sets Used in the Paper

| | InfoVis-VAST | Car Reviews | TV Reviews | 20News |
|---|---|---|---|---|
| #docs | 515 | 231 | 110 | 2,211 |
| #words | 5,935 | 3,142 | 1,624 | 23,604 |

method is applied in a straightforward manner. That is, when an intermediate output of t-SNE, which is a set of 2D representations of the documents, is available, the visualization thread updates the node-link diagram view. In addition, once an intermediate output of SS-NMF is generated, it causes the t-SNE module to restart because of the new input distance information due to the topic cluster membership changes. During this process, the latest output of t-SNE from the previous run of t-SNE is used as an initial value for the next run of t-SNE in order to avoid a significant change in the view.

While the view updates continuously based on the intermediate outputs of t-SNE as well as SS-NMF, the user can perform any supported user interactions, which will essentially restart SS-NMF using the newly created reference information. Similar to the restarting process of t-SNE, we set the initial value of the new run of SS-NMF as the latest output of the previous run of SS-NMF. Afterwards, the intermediate results of SS-NMF from iterations will immediately affect the t-SNE module and then the visualization module. In this manner, the SS-NMF module can be responsive due to user interactions in real-time.

### 5.3 Animated Visualization of Explicit Topic cluster Changes

During a sequence of user interactions, documents often change their topic cluster memberships. In UTOPIAN, such changes assign different colors to data points (representing new topic cluster indices), and their 2D coordinates computed by our modified t-SNE often change significantly. In order to preserve the user's mental map to track these changes [2], UTOPIAN visualizes their smooth transitions via animation. Furthermore, we explicitly encode the topic cluster changes of a data point by filling the left half of the point circle as the original cluster index and the right half as the new cluster index.[2]

### 5.4 Implementation Details

UTOPIAN is primarily implemented in JAVA for front-end UI's and rendering modules, which are mainly based on the FODAVA testbed system [8]. NetBeans Rich Client Platform and IDE[3] have been used for a flexible window management. The back-end computational modules of SS-NMF and t-SNE are written in MATLAB, and they interface with the front-end JAVA module via the 'matlabcontrol' library.[4] For animation effects, we have used the 'trident' library.[5]

## 6 Quantitative Analysis

We present quantitative comparisons between the LDA and the NMF algorithms from the two practical viewpoints: (1) *consistency from multiple runs* and (2) *empirical convergence* (as opposed to the algorithmic convergence). By the former, we mean how consistent results the algorithm generates among multiple runs while by the latter, we mean how fast the algorithm converges from a human's practical viewpoint.

In terms of the algorithm implementation, we use a widely-accepted LDA implementation called Mallet [28] the algorithm of which is based on a Gibbs sampling method [32]. For NMF, we have used one of the fastest and numerically reliable implementation based on an active set type of a least squares method [22].[6] For both methods, random initialization provided by each algorithm has been used.

### 6.1 Data Sets

We have chosen four document data sets: the InfoVis-VAST, the Car Reviews, the TV Reviews, and the 20News data sets. The InfoVis-VAST data set[7] is a collection of academic papers published in IEEE InfoVis (1995-2010) and VAST (2006-2010) conferences. The Car Reviews data set contains a set of reviews about 2009 Hyundai Genesis car collected from Edmunds.com, and the TV Reviews data set is another product review document set about a particular Samsung TV collected from Amazon.com. Finally, the 20News data set[8] is a collection of newsgroup documents composed of 20 topics. Notice that the 20News data set is relatively well clustered with pre-defined topic cluster labels. The size of these data sets are summarized in Table 2. After the data sets are encoded using a bag-of-words representation, we pre-processed them using tf-idf and unit $L_2$-norm normalization.

### 6.2 Consistency from Multiple Runs

In this experiment, we have run the LDA and the NMF algorithms multiple times on the four data sets. After obtaining ten sets of the topic modeling results from each algorithm, for each pair of the result sets within a particular algorithm, we have measured the number of the topic cluster memberships of individual documents that did not agree and averaged such a measure over 45 pairs out of ten sets. To this end, we have applied the Hungarian algorithm [24] to match the two independently generated cluster index sets.

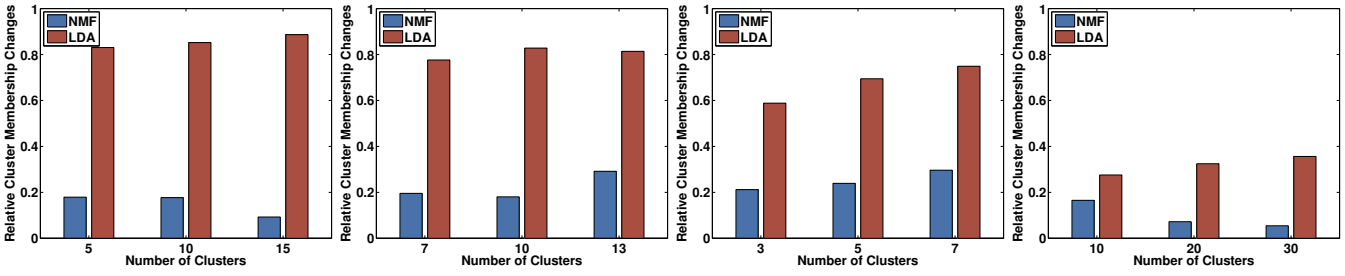
Fig. 3 shows the relative number of averaged topic cluster membership changes out of the total number of documents depending on the number of topics for each data set. In the first three data sets, it is shown that using NMF, about 10-25% of the entire documents changed their topic cluster memberships while using LDA, the corresponding value was much high, e.g., around 60-85%. It indicates that *the user will observe significant topic cluster membership changes each time running LDA whereas the changes are relatively minor in the case of NMF*.

Compared to the first three data sets, however, in the 20News data set, the number of topic cluster membership changes in LDA decreases to around 30%. NMF also performs better than the previous data sets, but the performance improvement is not significant compared to LDA. Interestingly, this observation tells us that LDA tends to give more consistent results for the data sets composed of clear topics such as the 20News data set. However, in many real-world data sets, the topics are often not clearly defined due to the presence of a significant amount of noise, which makes NMF more viable for analyzing the real-world data in a visual analytic environment.

Such inconsistent behaviors of LDA can also be shown by the topic summary that each algorithm generates. For example, Table 2 shows the topic summaries from the two result sets from the InfoVis-VAST data set. In the case of NMF, the topic summaries are shown to be exactly the same, but LDA varies significantly for some topics. For example, topic 2 is shown to have completely different topic summaries ('knowledge, edge' vs. 'analysts, scatterplot'), and so is topic 4 ('social, tree' vs. 'text, document').

These experimental results implies important practical concerns about LDA when used in visual analytics domains since the user cannot assure that the LDA result at hand is the best one for his/her analysis. For example, one LDA result may fail to reveal a particular topic of interest while some other ones can. To overcome this issue, the user could run LDA multiple times and see if any interesting differences arise, but this process could become time-consuming due to a significant running time of LDA and a nontrivial task of comparing between different LDA results.

[2] For more details, please refer to the accompanying video.
[3] http://netbeans.org/features/platform/index.html
[4] https://code.google.com/p/matlabcontrol
[5] https://kenai.com/projects/trident/pages/Home
[6] http://www.cc.gatech.edu/~hpark/nmfsoftware.php.
[7] http://www.cc.gatech.edu/gvu/ii/jigsaw/datafiles.html
[8] http://qwone.com/~jason/20Newsgroups

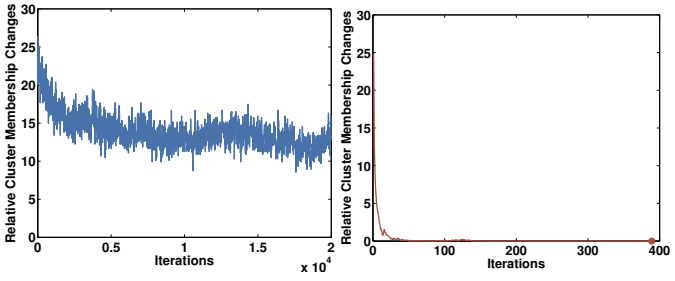



(a) The InfoVis-VAST data set (b) The Car Reviews data set (c) The TV Review data set (d) The 20News data set

Fig. 3. The relative topic cluster membership changes with respect to the total number of documents. The presented results are averaged values among the sets of cluster membership results from running each algorithm ten times with different numbers of clusters.

Table 3. Topic Summaries from the Two Runs of NMF and LDA

| NMF | | | | | | |
|---|---|---|---|---|---|---|
| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 |
| visualization,design | information,user | analysis,system | graph,layout | visual,analytics | data,sets | color,weaving |
| visualization,design | information,user | analysis,system | graph,layout | visual,analytics | data,sets | color,weaving |
| **LDA** | | | | | | |
| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 |
| documents,similarities | knowledge,edge | query,collaborative | social,tree | measures,multivariate | tree,animation | dimensions,treemap |
| documents,query | analysts,scatterplot | spatial,collaborative | text,documents | multidimensional,high | tree,aggregation | dimensions,treemap |




(a) LDA (b) NMF

Fig. 4. The relative topic cluster membership changes between the two consecutive iterations on the InfoVis-VAST data set using LDA and NMF. The number of topics is set to seven.

### 6.3 Empirical Convergence

Now, we have analyzed the convergence behavior of each algorithm using a practically meaningful measure to a human's perception rather than the algorithm's own objective or criterion values. As the measure, we have used the relative number of topic cluster membership changes between iterations out of the entire documents. In the case of NMF, we computed this measure until the NMF algorithm converges based on its own convergence criteria that are well designed from their theoretical and numerical analysis. In the case of the LDA algorithm implemented in Mallet, we did not find any convergence or stopping criteria other than a fixed number of iterations. By default, Mallet is set to run 2,000 iterations, but in our experiments, we have run 20,000 iterations, which we think of as a sufficiently large number of iterations.

Fig. 4 shows the relative number of averaged topic cluster membership changes between the two consecutive iterations for the InfoVis-VAST data set.[9] Surprisingly, even after numerous iterations (20,000), it is shown that around 10-15% of the total documents change their topic cluster memberships at every iteration, showing practically no convergence at all. This is partly because of the *Randomness* nature of the sampling-based LDA algorithm, which we briefly mentioned in Section . In other words, each iteration is mainly performed by drawing samples from a particular distribution, which is expected to converge or to be stable as iterations go on. However, the result indicates it is not the case.

This *Randomness* nature of the LDA algorithm basically gives the user no control over the algorithm process. To be specific, suppose the user obtained one LDA result from a particular run, but she wants to improve the result by slightly modifying an initial specification of the LDA algorithm, such as an initialization and a parameter of LDA. However, such a user intervention may yield a completely different result from what the user previously had since the sampling-based LDA algorithm is affected not only by an initial specification but also by the samples randomly generated at each iteration. In contrast, in most deterministic algorithms such as the NMF one we used, as long as the user fixes the initial specification, the result can be replicated, and thus the user can isolate and control the effects of his/her intervention while running an algorithm.

Now let us discuss about the behavior of the NMF algorithm shown in Fig. 4(b). When converged (at 389 iterations), NMF does not show such undesirable behaviors of LDA, giving a stable result with no topic cluster membership changes. More importantly, using NMF, the number of cluster membership changes decreases quite quickly to almost zero at early iterations, e.g., around 60 iterations out of 389 in total. It indicates that the user can get most information from NMF at a much shorter period of time than that required for a full convergence of the NMF algorithm. This behavior justifies the usefulness of the adopted iteration-wise visualization framework (Section 5.2) so that the user can immediately analyze the NMF result and perform various interactions in real-time.

### 6.4 Running Time

We will now briefly present the running time taken in the experiments shown in Fig. 4. While LDA has taken about 10 minutes for 20,000 iterations, NMF has taken 48 seconds until the convergence. In the case of LDA, if the number of iterations is set to a smaller value, the running time linearly decreases, but in general, we found that the running time of NMF is still much faster than LDA, which gives another practical advantage of NMF over LDA especially in the context of visual analytics.

## 7 Usage Scenarios

In this section, we present several usage scenarios showing the user interaction capabilities of UTOPIAN, which is based on SS-NMF (Section 4), for improving the topic modeling result in a user-driven manner.

[9] Due to a page limit, we omitted the results of the other data sets. The overall behaviors of them were similar to the InfoVis-VAST data set.

### 7.1 The TV Reviews Data

As shown in Fig. 5, we performed several user interactions to have better understanding about the data. First, we have initially run the NMF topic modeling with five topics. As shown in Fig. 5(a), the initial result from the NMF topic modeling reveals interesting topic clusters from the data set. For example, the topic labeled as 'delivery, service, amazon' mainly talks about the delivery service from 'Amazon.com.' Another one labeled as 'money, worth, spent' generally mentions that people are satisfied with the price, e.g., "Well worth the money spent!".

Now, we focused on the cluster labeled as 'television, excellent, product,' which would likely contain positive reviews about the product. However, the keyword such as 'television' is not much meaningful in this data set. Therefore, we performed *topic keyword refinement* for removing this keyword but instead increasing the weight of the keyword 'recommended' originally ranked as the 11-th keyword. As seen in Fig. 5(b), the resulting topic is now labeled as 'excellent, product, recommended' reflecting this refinement process.

More interestingly, we found one document that has moved from this topic cluster to another labeled as 'dvd, problem, sound.' After reading this review, we found that this document, which starts by saying "Do not buy this TV!!", indeed discussed mostly the negative features of the product. Now, we decide to study and understand the negative aspects of this product, and thus, we have performed *document-induced topic creation* with this document. As seen in Fig. 5(c), the newly created topic labeled as 'repair, problem, stopped' cluster has been created, and in this topic, we have found three new documents that mainly mention about the problematic issues such as the lengthy time taken during a warranty repair, a loud noise from speakers, pink- and green-colored dead pixels, a connectivity issue with a Verizon FIOS set-top box.

### 7.2 The Car Reviews Data

Using this data set, we describe the use cases of *keyword-induced topic creation* and *topic splitting*. Given the initial result shown in Fig. 6(a), an interesting topic cluster is the one labeled as 'problem, shift, gears' that could imply that this car largely has an issue with gear shifting. As reading a few documents in this topic cluster, we found that many reviews complain about the transmission/gear shifting system, e.g., "Rattling/grinding sound when driving at lower gears", "The transmission clunks.", "The Transmission shifting is a Nightmare and very embarrassing!".

Next, we wanted to see if there is any suspension-related problems, and thus we have performed *keyword-induced topic creation*, by using the keywords 'problem' and 'suspension' (black circles in Fig. 6(a)). As a result, a new topic labeled as 'suspension, problem, design' has now contained multiple documents (black circles in Fig. 6(c)) that mentions about the suspension issue, e.g., "Jittery suspension.", "Redesign the whole suspension system.", "Suspension ruins the whole car comfortability.".

In addition, we performed *topic splitting* on an unclear topic labeled as 'seats, mileage, passengers' (a black triangle in Fig. 6(a)) by manipulating the keywords as shown in Fig. 6(b). Due to this interaction, UTOPIAN properly split the clusters by isolating the reviews mentioning the gas mileage, e.g., "great mileage for a 3.8L", "I like the gas mileage of the v6. hwy 28 mpg and 23-24 mixed driving." vs. the features about the seat, e.g., "add programmable front passenger seat", "cooled seats".

### 7.3 The InfoVis-VAST Data

In this scenario, we utilize *topic merging*, *topic splitting*, and *keyword-induced topic creation* for the InfoVis-VAST data set. As can be seen in Fig. 7(a), the initial result computed by NMF is quite comprehensive, revealing an overview about the research topics in information visualization and visual analytics fields. For instance, the topic labeled as 'document, text, collections' is shown to be mainly about text visualization, and the one labeled as 'networks, traffic, social' is shown to be about social network visualization.

Among these clusters, several topic clusters such as the ones labeled as 'treemaps, layout, hierarchical' and 'trees, hierarchy, node' are shown to be similar, and thus we have merged these topics (black circles in Fig. 7(b)). On the other hand, we have split the topic cluster labeled as 'dimensions, multivariate, parallel' in order to further look into the specific research about 'dimension reduction' and that about 'cluster analysis.' After performing such interactions, we have found the topics have been properly merged/split as expected. For instance, the papers such as 'Interactive Visual Clustering of Large Collections of Trajectories' and 'ClusterSculptor: A Visual Analytics Tool for High-Dimensional Data' have been clustered to the 'cluster analysis' topic while the papers such as 'A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections' and 'Interactive Dimensionality Reduction through User-defined Combinations of Quality Metrics' have been clustered to the 'dimension reduction' topic. It should be noted that although we manipulated only the keywords 'dimension,' 'reduction,' 'cluster,' the relevant documents without these specific keywords have been properly clustered.

Next, we have focused on the cluster labeled as 'graph, layout, edge' (a black rectangle in Fig. 7(b)). We performed *keyword-induced topic creation* based on the keyword 'edge' to look into the research about 'edge' in the context of graph visualization. Interestingly, the result has shown the topic labeled as 'edge, bundled, adjacencies,' which implies that the edges in the graph are mainly used to represent the adjacencies and edge bundling is one of the main research topics. Although not reported, another keyword 'crossings' has been highly ranked as the fifth one, which also makes sense in that edge crossings are one of the main issues in graph visualization. The example papers from this topic cluster include 'Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data' and 'Edgelens: An Interactive Method for Managing Edge Congestion in Graphs.'

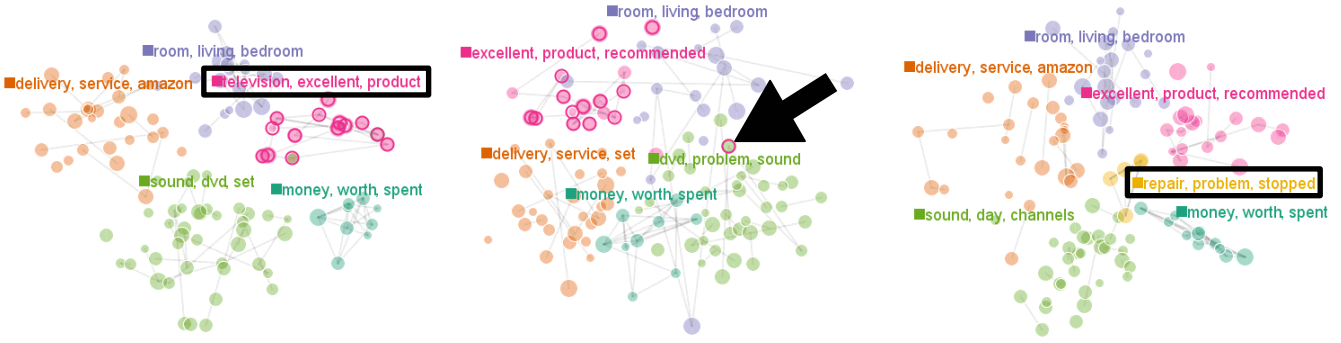
## 8 Conclusions and Future Work

In this work, we presented an NMF-based interactive visual topic modeling system called UTOPIAN. Compared to a widely-used topic modeling method, LDA, NMF has many practical advantages from the perspectives of *consistency from multiple runs* and *early empirical convergence*. In addition, NMF can incorporate the user input based on the semi-supervised formulation of it in an intuitive way. By utilizing the semi-supervised NMF, we provided five useful interaction capabilities, *topic keyword refinement*, *topic merging*, *topic splitting*, *document-induced topic creation*, and *keyword-induced topic creation*. Next, we presented several key advantages of UTOPIAN such as the modified t-SNE algorithm, the iteration-wise visualization of NMF and t-SNE for supporting various user interactions in real-time. Finally, we demonstrated the capabilities of UTOPIAN by flexibly applying the supported interactions in several real-world data sets.

Beyond document analysis, UTOPIAN can be flexibly extended in visual analytics for various other domains, such as bioinformatics, network analysis, etc., owing to the its easy interpretation and interaction capabilities. As our future work, we plan to extend UTOPIAN for dealing with streaming document data. In addition, instead of simple keyword summaries, we plan to strengthen summarization capabilities showing semantically meaningful phrases or representative sentences, which would give the user much more comprehensive understanding about the resulting topics. Finally, we plan to improve UTOPIAN for handling a large-scale data based on a parallelized distributed NMF algorithm.

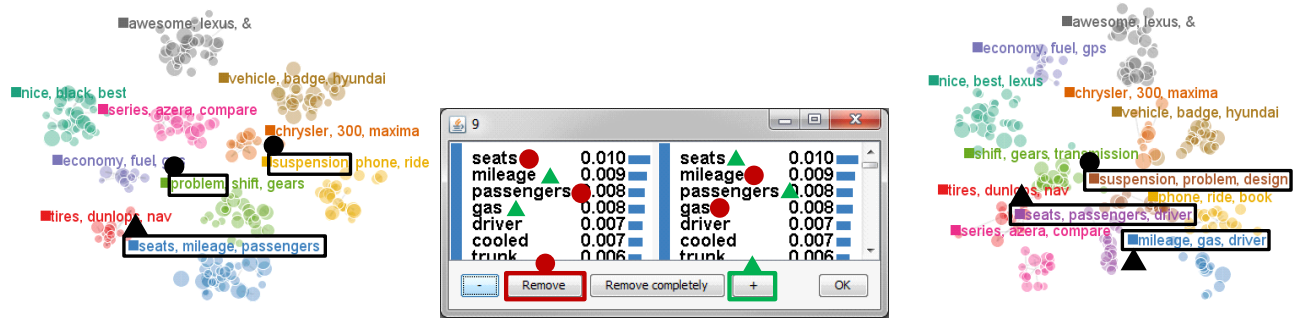
## 9 Acknowledgments

The work of these authors was supported in part by the National Science Foundation (NSF) grants CCF-0808863, IIS-1242304, and IIS-1231742, and the Defense Advanced Research Projects Agency (DARPA) XDATA program grant FA8750-12-2-0309. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF and DARPA. Finally, we would like to thank Juwon Drake and Jaewon Drake for their significant help in improving the manuscript and the video material.



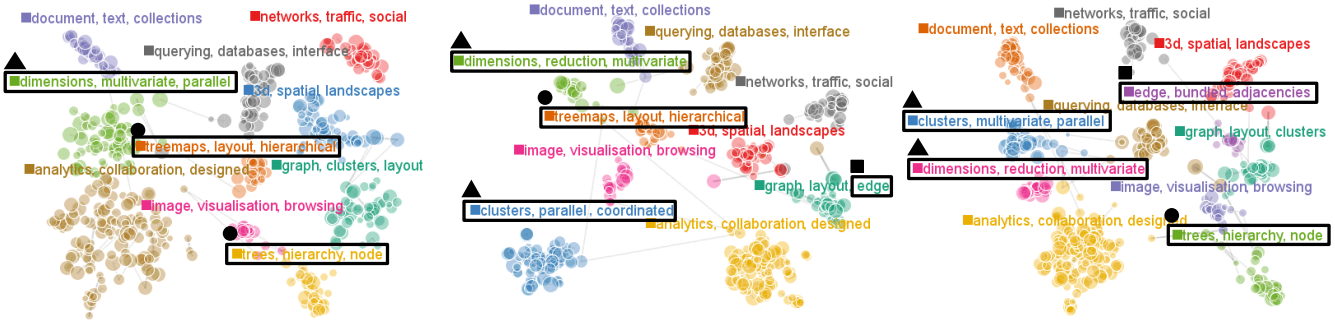

(a) The initial visualization (b) The visualization after *topic keyword refinement* (c) The visualization after *document-induced topic creation*

Fig. 5. The usage scenario with the TV Reviews data set. Given the initial visualization (a), we performed *topic keyword refinement* on the highlighted topic containing the term 'excellent' by removing the term 'TV' and by increasing the weights of the term 'recommended.' Due to this interaction, a single document (pointed by an arrow) has moved from this cluster to the other containing a keyword 'problem' (b). After reading it, this document is shown to mostly complain about the product. Now, we performed *document-induced topic creation* by using this document. As a result, three more documents that contain mostly negative reviews have joined this topic cluster, which is also reflected in the keyword summary containing 'repair' and 'stopped.'




(a) The initial visualization (b) The visualization after *topic keyword refinement* (c) The visualization after *keyword-induced topic creation* and *topic splitting*

Fig. 6. The usage scenario with the Car Reviews data set. Given the initial visualization (a), we have performed *keyword-induced topic creation* and *topic splitting*. For the former, in order to look into any suspension issues, we have chosen the keywords 'suspension' and 'problem' (black circles) for a newly created topic (a). For the latter, we have split the unclear topic labeled as 'seats, mileage, passengers' (a black triangle) to the two where we have excluded the keywords 'seats' and 'passengers' but increased the weights of 'mileage' and 'gas' in the left while doing the opposite in the right (b). The result shows the newly created topic cluster about 'suspension, problem' and also the two well separated clusters (c).




(a) The initial visualization (b) The visualization after *topic keyword refinement* (c) The visualization after *keyword-induced topic creation* and *topic splitting*

Fig. 7. The usage scenario with the InfoVis-VAST data set. Given the initial visualization (a), we performed *topic merging* and *topic splitting* and (b) and then *keyword-induced topic creation* (c). First, we merged the two topic clusters commonly dealing with hierarchical data (black circles) and split the topic cluster about 'multivariate data visualization' (a triangle) to the one about 'dimension reduction' and the other about 'clustering' by increasing the weights for the corresponding terms (b). Afterwards, we performed *keyword-induced topic creation* based on a keyword 'edge' in the cluster about 'graph visualization' (a rectangle). The final result reveals a newly created topic labeled as 'edge, bundled, adjacencies' (c), revealing the relevant sub-topics such as edge bundling, edge crossing, etc.

## REFERENCES


[1] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *Proc. the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, pages 1171–1177, 2011.

[2] D. Archambault, H. Purchase, and B. Pinaud. Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE Transactions on Visualization and Computer Graphics*, 17(4):539–552, 2011.

[3] S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond svd. In *Proc. IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1–10, 2012.

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

[5] K. Chen and L. Liu. ivibrate: Interactive visualization-based framework for clustering large datasets. *ACM Transactions on Information Systems (TOIS)*, 24(2):245–294, 2006.

[6] J. Choo, S. Bohn, and H. Park. Two-stage framework for visualization of clustered high dimensional data. In *Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 67–74, 2009.

[7] J. Choo, H. Lee, J. Kihm, and H. Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 27–34, 2010.

[8] J. Choo, H. Lee, Z. Liu, J. Stasko, and H. Park. An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data. In *Proc. SPIE 8654, Visualization and Data Analysis (VDA)*, pages 1–15, feb 2013.

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

[10] 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*, 17:2412–2421, 2011.

[11] 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.

[12] W. Dou, X. Wang, R. Chang, and W. Ribarsky. Paralleltopics: A probabilistic approach to exploring document collections. In *Proc. IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 231–240, 2011.

[13] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Proc. IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 93–102, 2012.

[14] S. M. Drucker, D. Fisher, and S. Basu. Helping users sort faster with adaptive machine learning recommendations. In *Proc. the 13th IFIP TC 13 International Conference on Human-computer Interaction (INTERACT) - Volume Part III*, pages 187–203, 2011.

[15] Z. Geng, R. S. Laramee, F. Loizides, and G. Buchanan. Visual analysis of document triage data. In *Proc. International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 151–163, 2011.

[16] B. Gretarsson, J. Odonovan, 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.

[17] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. the 22nd Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, pages 50–57, 1999.

[18] M. S. Hossain, P. K. R. Ojili, C. Grimm, R. Muller, L. T. Watson, and N. Ramakrishnan. Scatter/gather clustering: Flexibly incorporating user feedback to steer clustering results. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2829–2838, 2012.

[19] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):995–1006, 2004.

[20] Y. Hu, J. Boyd-Graber, and B. Satinoff. Interactive topic modeling. In *Proc. Association for Computational Linguistics*, pages 248–257, 2011.

[21] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[22] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.

[23] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[24] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

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

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

[27] A. Lopes, R. Pinho, F. Paulovich, and R. Minghim. Visual text mining using association rules. *Computers & Graphics*, 31(3):316–326, 2007.

[28] A. K. McCallum. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`, 2002.

[29] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre. Clustersculptor: A visual analytics tool for high-dimensional data. In *Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 75–82, 2007.

[30] P. Paatero and U. Tapper. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[31] F. Paulovich and R. Minghim. Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1229–1236, 2008.

[32] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proc. the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577, 2008.

[33] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.

[34] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[35] 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 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 153–162, 2010.

[36] J. A. Wise. The ecological approach to text visualization. *Journal of the American Society for Information Science*, 50(13):1224–1233, 1999.