HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies

Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky

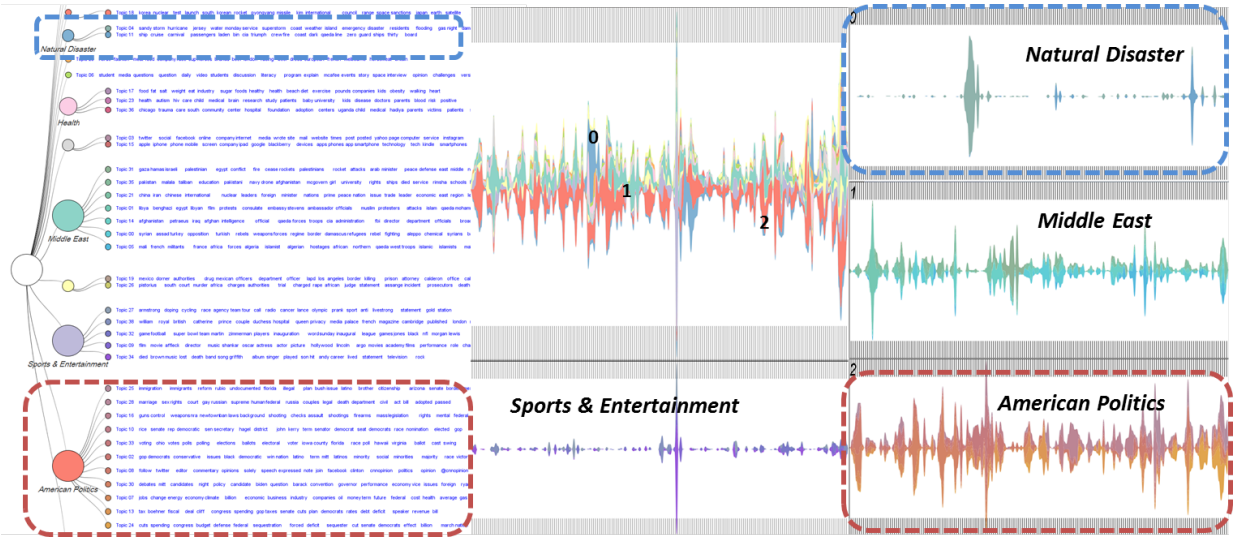

Fig. 1. Overview of the HierarchicalTopics system. The Hierarchical Topic structure is shown on the left in a tree visualization. The Hierarchical ThemeRiver view on the right presents the temporal pattern of topics in a hierarchical fashion. The dataset being visualized is the CNN news corpus. Topics are organized into 5 categories and annotations are attached to describe each news category. The corresponding categories in both view are outlined with same colors.

Abstract—Analyzing large textual collections has become increasingly challenging given the size of the data available and the rate that more data is being generated. Topic-based text summarization methods coupled with interactive visualizations have presented promising approaches to address the challenge of analyzing large text corpora. As the text corpora and vocabulary grow larger, more topics need to be generated in order to capture the meaningful latent themes and nuances in the corpora. However, it is difficult for most of current topic-based visualizations to represent large number of topics without being cluttered or illegible. To facilitate the representation and navigation of a large number of topics, we propose a visual analytics system - HierarchicalTopic (HT). HT integrates a computational algorithm, Topic Rose Tree, with an interactive visual interface. The Topic Rose Tree constructs a topic hierarchy based on a list of topics. The interactive visual interface is designed to present the topic content as well as temporal evolution of topics in a hierarchical fashion. User interactions are provided for users to make changes to the topic hierarchy based on their mental model of the topic space. To qualitatively evaluate HT, we present a case study that showcases how HierarchicalTopics aid expert users in making sense of a large number of topics and discovering interesting patterns of topic groups. We have also conducted a user study to quantitatively evaluate the effect of hierarchical topic structure. The study results reveal that the HT leads to faster identification of large number of relevant topics. We have also solicited user feedback during the experiments and incorporated some suggestions into the current version of HierarchicalTopics.

Index Terms—Hierarchical topic representation, topic modeling, visual analytics, rose tree

1 Introduction

Digital textural content is being generated at a daunting scale, much larger than we can ever comprehend. Vast amounts of content is accumulated from various sources, diverse populations, and different times and locations. For example, 1.35 million scholarly articles were published in 2006 alone [18]. With an average annual growth rate of 2.5% [30], research articles are currently being published at the pace of approximately 4400 titles per day. In the social media world, people are contributing to the accumulation at an even faster pace. By June 2012, Twitter is seeing 400 million tweets per day [31]. Meanwhile, 900 million active Facebook users have been busy sending 1 million messages every 20 minutes [28]. Today, part of the content (e.g, tens of thousands of different sites, Twitter, digitized books) is archived in the US Library of Congress with more than 300 terabytes in size, which keeps on growing [11].

It is generally agreed in government and industry that valuable but

- *Wenwen Dou is with University of North Carolina at Charlotte. E-mail: wdou1@uncc.edu.*
- *Li Yu is with University of North Carolina at Charlotte. E-mail: lyu8@uncc.edu.*
- *Xiaoyu Wang is with University of North Carolina at Charlotte. E-mail: xwang25@uncc.edu.*
- *Zhiqiang Ma is with University of North Carolina at Charlotte. E-mail: zma5@uncc.edu.*
- *William Ribarsky is with University of North Carolina at Charlotte. E-mail: ribarsky@uncc.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.

latent information is hidden in the vast amount of digital textual content. For instance, in scientific research, one of the crucial investigations is on the development of science. To this aim, researchers have created maps of science [25, 27] and evaluated the impact of science funding programs [14] by analyzing research publications and proposals. For emergency response agencies, sifting through massive amount of social media data could help them monitor and track the development of and response to natural disasters, as illustrated in the use of Twitter to reach victims from Hurricanes [35]. Last but not least, the emergence of numerous social media startups shows that profitable marketing and business analytics insights that can be extracted from such content. To extract insights and make sense of large amounts of textual data, efficient text summarization is therefore much needed.

In this regard, topic models have been considered as the state-of-the-art statistical methods to extract meaningful topics/themes for summaization. Although powerful, topic models do not provide meanings and interpretation; human must be involved [7]. To enhance the interpretations of topical results, visual text analytics researchers have designed algorithms and visual representations that make the probabilistic topic results legible and exploratory to a broader audience [8, 9, 10, 14, 15, 26, 34]. Examples of the utility of these topic-based visualization interfaces include the analysis of social media users based on the content they generated [22], depiction of the temporal evolution of topics [14, 26], and identification of interesting events from news and social media streams [8, 15]. Many of these topic-based visualization systems have been studied through use cases and regarded powerful in aiding text analysis processes.

However, current visual text analytics systems have limitations. In contrast to the common practice of extracting hundreds of topics from large document corpora in the topic model community [2, 4, 21, 29, 32], current systems usually only manage to effectively represent a small number of topics. As more textual data becoming available, the number of necessary topics for interpretable text summarization will grow inevitably. Only extracting a small number of topics, therefore, won't capture the nuances in the corpora. As the number of topics increase, sifting through and comprehending all the topics becomes a time-consuming and laborious task, which will be further hampered by the visual clutter introduced when displaying the temporal evolution of hundreds of topics with no organization.

In particular, three challenges must be met to effectively analyze document collections that are summarized by large number of topics:

1. **How to organize the topics to facilitate the navigation and analysis within the topic space?** Without organization, sifting through a hundred topics with each topic consisting of 20 or more keywords could be intimidating. One example that highlights the problem is that when developing the NSF Portfolio Explorer, it took days for a researcher to manually examine a thousand topics to select 30 topics for further analysis and visualization [12]. Since certain topics are closer in meaning than others, organizing semantically similar topics into topic groups will ease the navigation in the topic space. Having an automated classification of topics could potentially jumpstart the analysis of text collections based on large number of topics, however, the automated classification may not always conform to individual users' mental model of the topics space.

2. **How to visually convey and permit user interactions with the organized topic results so that users can classify the topics based on their interests?** It is essential to place users in the center of the topic analysis process, allowing users to leverage and modify the topic classification results. For example, when analyzing a news corpus, a user may want to organize the topics into a hierarchical structure through first categorizing the news topics into either domestic or foreign news. In addition, for domestic news topics, the user may want to further divide the topics into groups such as politics, sports, entertainment, etc. Similarly, when analyzing topics from Twitter streams, a business analyst may be interested in grouping all topics related to sales and customer services and further divide them into more refined categories. Therefore, intuitive topic visualizations and user interactions are needed to support the analysis and modification from an initial topic organization provided by an automated algorithm.

3. **How to modify existing visual metaphors to accommodate the organization of a large number of topics?** After a user has identified a desirable hierarchical topic structure, the third challenge lies in tailoring existing visual representations. Visualizing temporal evolution of topics has been considered essential to understanding various domains (e.g. scientific fields, breaking news, etc.) over time. However, ThemeRiver [16] and stack graph that are commonly used to present the temporal trends of the topics do not convey hierarchical information. To enable the analysis and comparison of temporal behavior of topic and topic groups, it is essential to extend the current visual metaphors to incorporate hierarchical structure of topics.

To tackle the three challenges, we propose HierarchicalTopics (HT), a visual analytics system [1] that supports scalable exploration and analysis of document corpora based on a large number of topics. HierarchicalTopics *addresses the first challenge* by integrating a novel algorithm that automatically classifies topics into a hierarchical structure. Through joining similar topics into the same group, the new organization of topics provides scalable representation and navigation in the topic space. HierarchicalTopics further incorporates visual representations and interactions that embrace the hierarchical organization of the topics, and enables the users to depict the temporal evolution of topics or topic groups. In addition, user interactions are provided in HT to *address the second challenge*. Along with the visual representations of the topic hierarchy, HT allows users to modify and update the automatically computed topic groups. It therefore supports the customization of the visualizations based on the users' analytical interests. *To address the third challenge*, a new Hierarchical ThemeRiver has been designed to accommodate the hierarchical organization of the topics. The Hierarchical ThemeRiver eases the exploration of temporal behaviors of topic groups, and enables the comparison of topic groups on a temporal dimension. Through tight coordination between the visualizations of topic hierarchy and hierarchical temporal trends, we intend to provide an inviting interface that supports making sense of large document collections via navigating through large number of topics and their temporal evolution.

We have assessed the HT through both qualitative and quantitative evaluations. To evaluate the system in a qualitative manner, we present a case study in which an expert user performed in depth analysis on a collection of 11,961 NSF awarded proposal abstracts. To evaluate HT in a quantitative fashion, an 18-participant user experiment is conducted to compare the HierarchicalTopics system to a non-hierarchical representation based on a CNN news corpus that contains 2453 recent news articles. The experiment results reveal that the hierarchical topic visualization leads to faster identification of a large number of relevant topics. Constructive user comments were also collected during the experiment. After the user study, some suggestions on improving the visualization and interactions from the participants have been incorporated into the current version of the HierarchicalTopic system.

The rest of the paper is structured as follows: we introduce the previous work that inspired the design of HierarchicalTopics in Section 2. Section 3 focuses on introducing the HierarchicalTopics, including its system architecture and interactions. We present a case study in Section 4, followed by descriptions of a user study in Section 5.

2 Related Work

Two lines of work inspire the design of HierarchicalTopics, namely topic models and topic-based visualizations.

2.1 Topic Models

Topic models can be effective tools for text summarization and statistical analysis of document collections [2]. The number of topics

[1] A video of the HierarchicalTopics can be found at `http://youtu.be/Vi1FP5kAbOU`.

needed is typically determined by the size of the text corpora. The larger the size the more topics are preferred to ensure topic comprehension and human interpretability, typically tens of thousands of articles will require topics in the scale of hundreds. Specifically, one school of topic models is based on a human-defined number of topics. Researchers and practitioners usually generate a large number of topics to capture the themes that pervade the text collection as well as the nuances. For instance, in the experiment of evaluating the collaborative topic model [32], the authors extracted 200 topics from a paper-abstract collection with 16,980 articles and a vocabulary size of 8000. In other non-parametric Bayesian topic models, such as the hierarchical Dirichlet process (HDP) [29] and the discrete infinite logistic normal distribution (DILN) [21], the number of topics is determined by the model. However, it is evidenced that such algorithmically generated number of topics is typical rather large. For example, in the experiment evaluating DILN, the model produced 50 to 100 topics given a fairly small dataset with only 3000 to 5000 news articles.

Such large number of topics creates challenges to human interpretations and the sense-making process. Much research has been focused on revealing the correlations between latent topics and organizing topics into more human interpretable structures. Work in this area aims to facilitate the navigation through the topic space and enables the discovery of documents exhibiting similar topics. While most of the existing topic models do not explicitly model correlations between topics, a few exceptions have directly accounted for relationships between latent topic themes. For example, both correlated topic model (CTM) [4] and DILN [21] have demonstrated better predictive performance and have uncovered interesting descriptive statistics for facilitating browsing and search. Although the topic correlations have been modeled, it is still difficult for users to take advantage of the descriptive statistical relationship of topics without an effective organization and visual representation of the topics.

Many researchers consider that organizing topics into a hierarchical structure presents a scalable solution to improve human-interpretability of topic. To this aim, Blei et al. have proposed a hierarchical topic model (hLDA) that learns topic hierarchies from data to accommodate a large number of topics [3]. The hLDA is a flexible, general model for extracting topic hierarchies that naturally accommodates growing data collections. However, the topic hierarchies hLDA produced are rather rigid since the depth of such hierarchies is predefined and fixed throughout the modeling process. In addition, the higher level topics generated by hLDA usually consist of stopwords, therefore less meaningful for human users.

In order to leverage the scalable hierarchical structure without enforcing rigid restrictions on the topic models, we developed an algorithm, Topic Rose Tree, to construct a multilevel hierarchical structure with any given number of generated topics. Together with interactive visualizations, our HierarchicalTopics system enables users to explore and iteratively update the topic hierarchy. Our system aims to improve human-interpretability by enabling users to tailor the hierarchical topic results to their analytical interests or mental models of the topic space.

2.2 Visualization based on Topic Models

The power of topic models in summarizing and organizing large text corpora has been widely recognized in the visualization community. A good number of visualization systems have been developed based on topic models for users to comprehend document collections.

As one of the pioneer visual text analysis systems, TIARA [34] combined topic models and interactive visualization to help users explore and analyze large collections of text. Specifically, TIARA utilized a stack graph metaphor to represent temporal change of topics over time. Similarly, another system ParallelTopics was also developed to depict both temporal changes of topics using ThemeRiver and the characteristics of documents based on their topic proportions via Parallel Coordinates [14]. Since temporal evolution of the topics has been considered one of the most useful features of the topic-based visualizations, researchers have extended a great deal in this direction. TextFlow [13] presented a novel way to visualize topic birth, death, and merge that signify critical events. In a similar vein of identifying events, LeadLine [14] applied event detection methods to detect "bursts" from topic streams and further associate such bursts with people and locations to construct meaningful events. Furthermore, Chae et al. proposed a visual analytics approach that supports the analysis of abnormal events detected from topic time series [8]. Instead of representing and analyzing topics along the temporal dimension, Lee et al. proposed a visual analytics system for document clustering based on topic modeling [19]. Users could guide the clustering process through adjusting term weights in the topics.

These topic-based systems have demonstrate the effectiveness of combining topic models with interactive visualizations in facilitating analysis of text corpora. As indicated in in most of their reported case studies, however, these systems only dealt with a fairly small number of topics. This is quite contrary to the common practice in the topic modeling community, where a lot more topics are generated for a text collection of similar size (Section 2.1).While a greater number of topics will inevitably introduce visual clutter and legibility issue to the visualization systems, limiting the topic number may also hamper users' ability in comprehending the text collection.

Therefore, more scalable approaches to organizing the topics and visual representations based on the topics are much needed to support real-world challenges of analyzing large text corpora. To meet this need, HierarchicalTopics provides a scalable solution that allows iterative analysis of document collections with a large number of topics and further supports the exploration of temporal evolution of those topics in a hierarchical fashion.

3 HierarchicalTopics

3.1 System Pipeline

As illustrated in the overall system architecture in Figure 2, HierarchicalTopics is a user-centered analysis system that integrates computational methods with interactive visualizations. HT systematically incorporates both online and offline computations and utilizes scalable infrastructures described in [33], including MapReduce and Parallel Processing. There are four key processing stages in the HT architecture including two offline computation modules (e.g., Data collection, and preprocessing and Parallel Topic Modeling) and two online components (e.g., Topic Rose Tree and Hierarchical Visualizations).

In particular, HT accommodates digital text content from various sources including social media, research publications, news, etc. Once the data is collected, it is streamlined into HT's data cleaning and preprocessing step, as shown in Figure 2A. In this process, HT first unifies the formats of input data and converts certain documents (PDFs) to proper topic-model-readable text files. It then prepares the documents for parallel topic models by removing stopwords and emojis.

The cleansed data then goes through the topic modeling stage (Figure 2B), which extract topics from the document collection. It is worth noting that the choice of the topic model component in HT is rather flexible. The architecture of HT is set to utilize a variety of topic models and can leverage their unique strengths such as interpretability [7], convenience of non-parametric models [21, 29], and accounting for additional metadata [23, 24], etc. As reported in paper, HT has successfully incorporated both the vanilla LDA [5] and the Author Topic Model (ATM) [24] to handle the natures of different text corpora.

After the first two stages are accomplished offline, the rest of the computation and visualization are computed online. The Topic Rose Tree (TRT) shown in Figure 2C organizes the probabilistic topic results into a hierarchical structure, as detailed in next section. Based on the hierarchical topic organization, two coordinated interactive visualizations (Figure 2D) are designed to present and support interactive analysis of topics and temporal evolution of the topics.

The TRT and the visualizations are closely coupled through the user interactions provided by the HierarchicalTopics system. In particular, the three essential operations in the TRT algorithm (e.g., join, absorb, and collapse) are directly incorporated in the visualizations and interactions. Through direct visual manipulations, HT allows the users to perform the same operations to modify the initial topic hierarchies and iteratively derive the most interpretable topics groups based on their analytic interest.

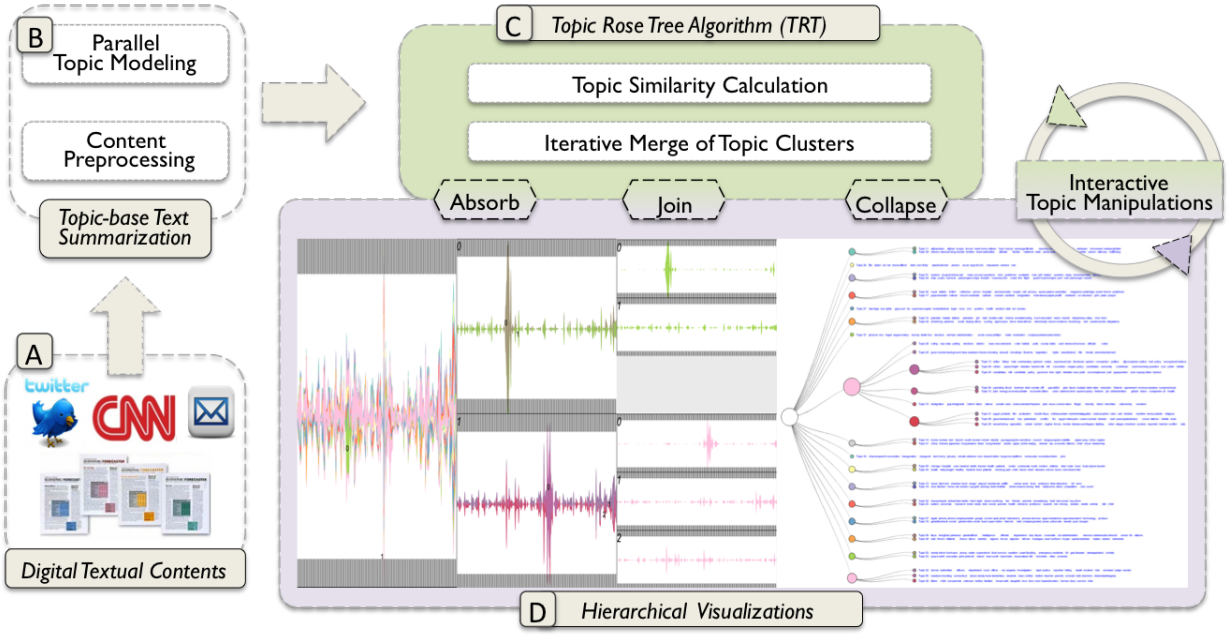

Fig. 2. System Architecture of HierarchicalTopics. Starting from bottom left, textual data is first harvested (A). The data then goes through a preprocessing stage before entering the topic model component (B). These two steps are completed offline. The resulting statistics from topic models then serve as input to the Topic Rose Tree (C), which constructs a hierarchy given a list of topics. The topic hierarchy is then visualized in the interactive visual interface (D) for users to analyze the topics and temporal trends in a hierarchical fashion to derive understanding of the text collection.

In the rest of this section, we will focus on presenting details of the online components of HierarchicalTopics.

3.2 Topic Rose Tree

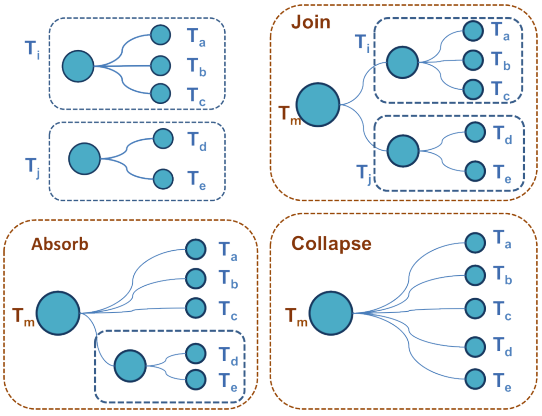

Fig. 3. The three essential operations of our Topic Rose Tree algorithm.

Our goal in designing the Topic Rose Tree is to support scalable visual representation and exploration. TRT is an automated method that can meaningfully organize a list of topics into a hierarchical structure. Its core algorithm is built upon key concepts from the Bayesian Rose Tree (BRT), which constructs a hierarchy using hierarchical clustering methods [6]. Compared to previous hierarchical clustering methods that limit discoverable hierarchies only to those with binary branching structures, BRT produces trees with arbitrary branching structure at each node, known as rose trees [6]. We consider such characteristic more natural in organizing topics, since any number of topics could be similar and should be grouped into one partition in a hierarchical structure. The essence of generating a rose tree is support of the three operations, namely join, absorb, and collapse (shown in Figure 3).

Unfortunately, simply borrowing BRT and directly applying it to topic models is unfit based on our experiments. This is primarily caused by the large number of features (words in the vocabulary) from topic models. In addition to the vocabulary size of a text corpus, which is usually in the thousands, the binarized matrix of topic distributions over the vocabulary is extremely sparse, causing problems for calculating the marginal probability of the topic groups in a tree.

Therefore, we developed TRT, an algorithm that built upon the three operations to construct hierarchies specifically from topic modeling results. TRT is a one-pass, bottom up method which initializes each topic in its own cluster and iteratively merges pairs of clusters. To construct the hierarchical structure, we first compute the similarity between any pair of clusters (topics/topic groups). TRT then merges the most similar clusters using one of the three operations. In this process, the Hellinger distance, which is a symmetric measure of the similarity between two probability distributions, is used to calculate the similarity of a pair of clusters. Intuitively, topics or topic groups that share similar distributions over the vocabulary yield lower distance. To construct the hierarchy, the most similar topic (group) clusters will be merged at each step.

In particular, each topic from the topic modeling results is represented as a probabilistic distribution over the entire vocabulary given a text collection, denoted by $X_{i,v}$, with i representing the ith topic and v representing the vocabulary of size N. To represent the probabilistic distribution of a node that contains multiple topics (children), we simply compute an average of all distributions of the children's. Details of the TRT are shown in Algorithm 1.

The complexity of the topic rose tree is the same as the BRT algorithm. First, the distance for every pair of data items needs to be computed-there are $O(n^2)$ such pairs. Second, these pairs must be sorted in order to find the smallest distance requiring $O(n^2 \log n)$ computational complexity.

To showcase how the topic rose tree algorithm could group similar topics together, Figure 4 shows a partial result from the initial grouping. In this case, we used the 2011 VAST mini challenge 1 microblog data, which contains an embedded scenario of an epidemic spread. This data is good for qualitatively evaluating the algorithm since we

Algorithm 1 Topic Rose Tree

Input: Data $\mathbf{D} = \{\mathbf{X_i}, \boldsymbol{v}\}, \boldsymbol{i} = 1,2,...n$; $\boldsymbol{v}$ is the vocabulary of the corpus
Output Topic rose tree T_{n+1}, a hierarchical structure with all topics
Initialize: $\mathbf{T_i} = \{\mathbf{X_i}, \boldsymbol{v}\}, \boldsymbol{i} = 1,2,...n$
Steps:

Denote c as cluster count
while $c > 1$ **do**
  for each pair of trees Ti and Tj **do**
    Calculate cost D(i,j) for 3 operations (join, absorb, or collapse):
    $D(i,j) = 1/2 * \sum_{v=1}^{N} (\sqrt{t_{i,v}} - \sqrt{t_{j,v}})^2$, $t_{i,v}$ denotes the probability distribution of tree node T_i over the vocabulary of size N
    Find operation m which yields lowest cost for T_i and T_j
    Merge T_i and T_j into T_m using operation m
    Delete T_i and T_j, $c = c - 1$
  end for
end while

expect similar topics regarding the epidemic spread should be grouped together. The topic group shown in Figure 4 (top) contains three topics highlighting the flu-like symptoms for the first two days of the epidemic (each tick on the x axis denotes a day). Another topic shown in Figure 4 (bottom) highlights evolved symptoms such as pneumonia for the third day of the epidemic. Note that since the words that were tweeted to describe the symptoms have changed a great deal, the topic rose tree did not put topic 31 into the first topic group. However, combining with the temporal patterns, one can identify when the epidemic spread started, and how the symptoms evolved over time. This example illustrates that the topic rose tree is able to group similar topics together, and the result is very much interpretable by human users.

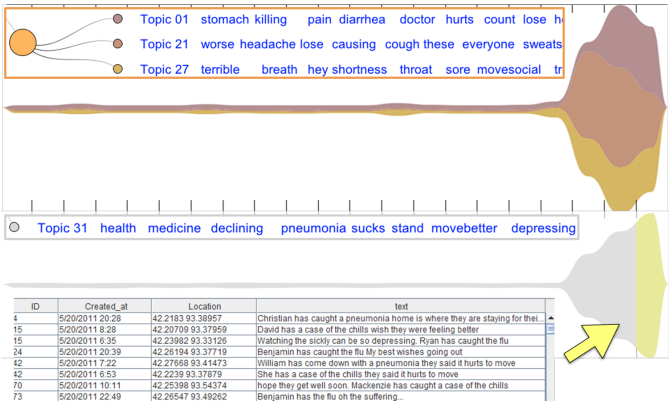

Fig. 4. An example showcases the capability of TRT grouping to group topics together. The top three topics (grouped by TRT) describe all flu-related symptoms on the first two days of the disease outbreak. The bottom topic (in grey) was not grouped into the first group by TRT since it describes different symptoms on the third day. Tweets related to certain topics are shown in a detailed view upon selection.

3.3 Visual Components

After applying the Topic Rose Tree to the topic modeling results, a hierarchical topic organization is generated. To facilitate the topic analysis of the text collection, we present a visual interface that is tailored to the hierarchical organization of topics. The visual interface consists of two coordinated views, namely Hierarchical Topic View and Hierarchical ThemeRiver. The two views are coordinated through user interactions with a focus on correlating the hierarchical information.

3.3.1 Hierarchical Topic view: Depicting topics in a hierarchical fashion

While TRT computationally alleviates the topic organization issue, the Hierarchical Topic view is designed to visually address ***Challenge 1*** by presenting the topic contents in a hierarchical fashion. Such representation not only offers a scalable solution as it allows the number of topics to accrue, but also supports better navigation by grouping similar topics together. Figure 1 shows the Hierarchical Topic view with 40 topics extracted from the CNN news corpus. To provide user a familiar visual environment, we adopt straightforward tree visual representation. In this view, each leaf node represents a topic, while the non-leaf nodes denote topic groups. The first node on the left is the root of the topic hierarchy, with the rose tree spanning from left to right. The content of each topic (in the form of a group of keywords) is presented to the right of each leaf node. The size of the node is drawn proportionally to its number of children (shown in figure 1).

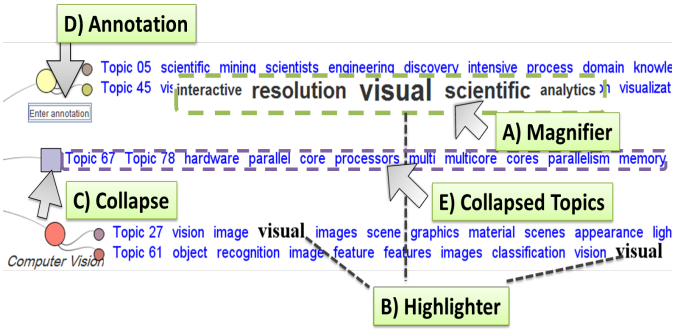

Fig. 5. Interactions provided by the Hierarchical Topic view. A) Magnifier: enlarges keywords near mouse cursor. B) Highlighter: highlight all occurrences of a selected keyword. C) Node collapsing: details of the collapsed children nodes are no longer shown. The shape of the node turns rectangular when collapsed. D) Annotation: allows users to enter annotation. E) Collapsed topic: keywords showing a summary of the two topics being collapsed.

User interactions. The Hierarchical Topic view provides a set of user interactions for effective exploration and navigation through large numbers of topics. In addition to standard panning and zooming, this view employs both an on-demand magnifier and highlighter to facilitate the examination of the topic contents, as shown in figure 5 A and B. The magnifier is designed to help users to better read the topic keywords through enlarging the font near the mouse cursor, while the highlighter aims to reveal the associations between topics by highlighting all occurrences of a certain keyword in the other topics. To further help users concentrate on the topics of interests, the Hierarchical Topic view supports interactive collapsing and expanding topic groups, shown in the square node in Figure 5C. Keywords for topics being collapsed into the same group are shown in (Figure 5E). More importantly, the Hierarchical Topic view allows users to annotate on the nodes to attach semantic meanings to topic groups (Figure 5D).

Interactive modification of the topic hierarchies. In addition to facilitating topic exploration, the Hierarchical Topic view aims to provide an intuitive way to visually classify the topics based on users' interest. In the process of analyzing a text corpus, only human users can attach semantics to the topics and provide meaningful yet sometimes subjective groupings. Therefore, it is essential to allow users to interactively modify the rose tree based on their analytical interests.

To permit such modification, the three operations that are used to construct the hierarchy in the topic rose tree algorithm are supported intuitively through drag-n-drop in the Hierarchical Topic view. As shown in Figure 6, dragging one leaf node into another constitutes the "join" operation. Drag-and-dropping any non-leaf node into another is considered as performing the "absorb" operation, while dragging multiple nodes into another node is interpreted as the "collapse" operation.

As observed in both the case study and user experiments (Section 4 and 5), the ability to iteratively refine and manipulate topic groups has demonstrated significant utility when analyzing text collections. Especially when HierarchicalTopics embodies the above three essential operations into intuitive mouse interactions, it creates a flexible text

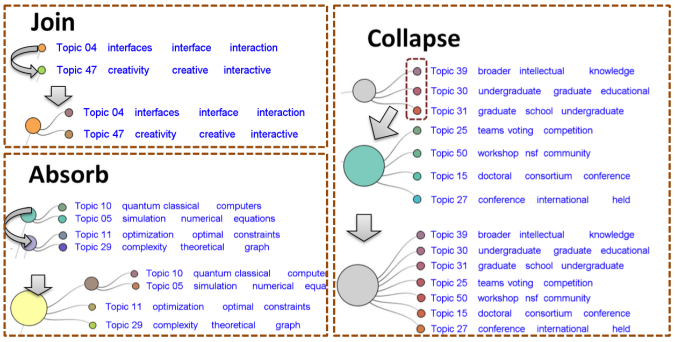

Fig. 6. Three operations supported to modify the topic hierarchy through user interactions.

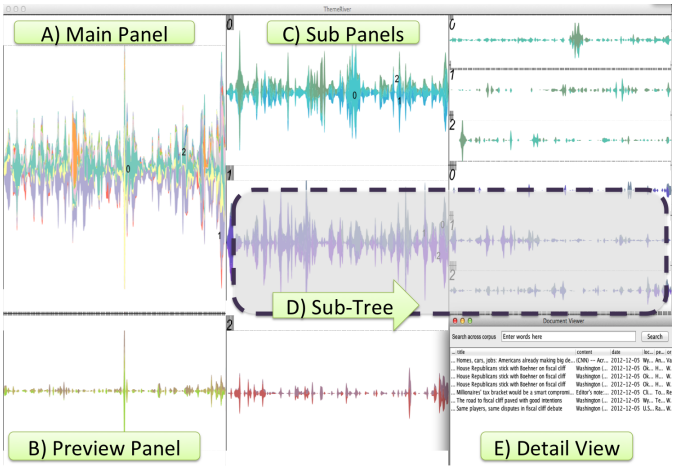

Fig. 7. Overview of the Hierarchical ThemeRiver. The dashed rectangle, in component D, highlights a sub tree created upon user interaction to view temporal patterns of child nodes.

analytics environment for users to categorize, modify, and update topics and topics groups. For example, as illustrated in Figure 1, participants in our user study have used these three operations to effectively group topics into five news categories based on the initial TRT hierarchy. In addition, the annotation interaction in HT view permits the users to attach semantic interpretations of the topic groups, and further helps them to connect the dots of a large number of topics. Many of our participants agreed that such user interactions served as a potential solution to the ***Challenge 2*** (see Section 1).

In summary, the Hierarchical Topic view provides both a visual representation of the topic hierarchy and a set of user interactions to serve as the first step to effectively analyze text collections.

3.3.2 Hierarchical ThemeRiver: Representing the temporal trends of topic groups

In addition to visually representing the topics which serve as a summarization of the document collection, visualizing the temporal evolution of the topics brings a unique contribution; it permits the discovery of the rise and fall of different topic themes, as well as identifying possible critical events [13, 15].

To this aim, we extend the widely adopted temporal visualization, ThemeRiver [16], to further incorporate hierarchical information. Our goal in designing the Hierarchical ThemeRiver is to provide users the ability to analyze and compare temporal behaviors of topic and topic groups, which address the core issue in ***Challenge 3*** (Section 1).

As illustrated in Figure 7, the Hierarchical ThemeRiver starts with the main panel (Figure 7A), where the temporal evolutions of the highest hierarchy (children of the root node) are shown; the height of each ribbon is calculated by summing the height of its leaf nodes. Once a ribbon is hovered, a preview of the temporal evolution of the child nodes will be shown in the preview panel (Figure 7B). The panels support interactively examination of the overall temporal trends of a text corpus as well as individual topic groups.

An elastic-panel structure is built into the view to enable the users' comparison of multiple topic groups. To compare different topic groups, a user can start by selecting a topic ribbon in the main panel; such interaction will create a sub panel (Figure 7C) showing the next level of hierarchy of the currently selected node. Multiple selections can be made to view the detailed temporal evolution of different topic groups, thus enabling the comparison and association of temporal patterns. Note that sub panels are always expanded to the right of the current selection, creating a coherent look and feel of the layout as in the Hierarchical Topic view.

Color assignment. To assist user exploration as well as to keep a smooth transition between panels, we have carefully chosen 12 perceptively coherent colors for the Hierarchical ThemeRiver view. This is done in an experimental fashion using the "i want hue" system [20], with the k-Means clustering and light background option. In the Hierarchical ThemeRiver view, the 12 distinct colors are first assigned to the topic ribbons in the main panel (Figure 7A). The child ribbons of each selected parent ribbon get colors of the same hue, but with varying luminance and chorma, as shown in Figure 7C. The same color scheme is also used in our Hierarchical Topic view to provide a coherent visual cue that helps correlating the two different representations of the same topic or topic groups.

Temporal selection and details on demand. To permit the examination of documents of interests, details of the text content are shown upon selection. In any panel within the Hierarchical ThemeRiver view, a user can enable the "time column" mode and interactively select a subset of documents published in a certain time period. By doing so, a detail view (Figure 7 E) will be shown to help the user validate the temporal patterns and understand its cause. During the user study, for example, this operation was demonstrated useful in examining the contributing posts to a topic burst pattern.

In summary, the Hierarchical ThemeRiver view is tailored to represent temporal patterns of topic and topics groups in a hierarchical manner. The incorporation of hierarchical information is mainly achieved through user interactions and in a way that is coherent to the Hierarchical Topic view representation.

3.3.3 View Coordination

Both views in the HierarchicalTopics system are tightly coordinated. On the one hand, selecting a node in the Hierarchical Topic view would highlight a corresponding temporal panel in the Hierarchical ThemeRiver view. This helps users to examine the temporal evolution of the selected topic group. On the other hand, selecting a ribbon in the temporal view will highlight the corresponding node and its path in the topic view. More importantly, once the hierarchy is modified through user interactions in the topic view, the temporal view will also be updated accordingly to reflect the new hierarchical structure.

In summary, the HierarchicalTopics system presents both topic information and temporal evolution of the topics in a hierarchical fashion. This system is designed to aid the exploration of topic content and temporal trends of topic groups through a set of user interactions. In addition, our system allows users to iteratively modify, define, and annotate topic groups based on their interpretation. The HierarchicalTopics provides a flexible visual analytics environment that tightly integrates computational methods with interactive visualizations for analysis of large document collections.

4 Case Study

To qualitatively access the utility of HierarchicalTopics in facilitating the analysis of text corpora with large number of topics, we recruited a senior researcher whose research interests covers HCI and Information Retrieval. This case study is set up for him to explore a collection of NSF awarded proposal abstracts to identify interesting research trends in his research domains. Eighty topics were extracted from 11,961 proposal abstracts funded by all three divisions (IIS, CCF, CNS) in

the CISE (Computer and Information Science and Engineering) directorate from 2005 to 2012.

4.1 Depicting temporal portfolio of NSF programs

Using the Hierarchical Topic view, the researcher started by visually browsing all hierarchical topic groups that are produced by the TRT algorithm. He quickly identified a few topics of interest and interactively merged them into topic groups that fits his analytic goal. The result of his customized grouping and corresponding annotation is shown in the first column in Figure 9. Specifically, two groups of topics are created through the "join" and "collapse" interactions, "*HCI*" and "*Information Retrieval and Data Mining (IR)*".

With the exploration scope narrowed down to these two topic groups, the user wanted to identify and compare the trends in research funding for individual group over the years. Therefore, he turned to the Hierarchical ThemeRiver view and selected the two topic groups so that their research funding trends can be examined and compared.

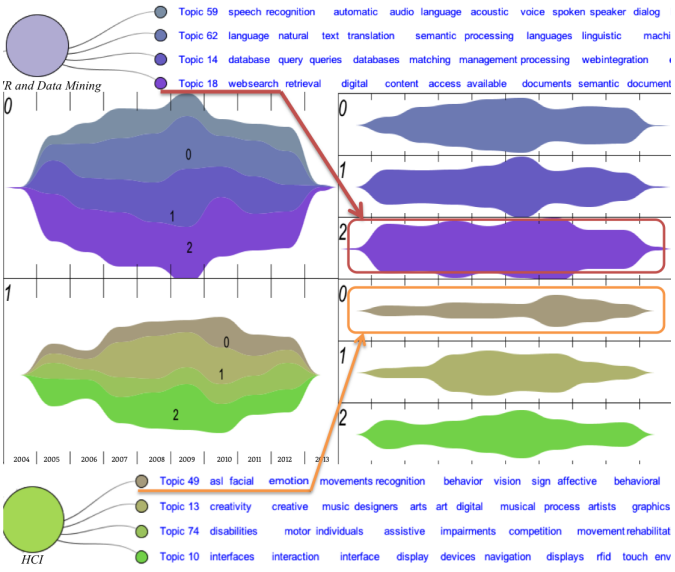

Fig. 9. Case Study: Examination of topic groups of interest. Top (with a purple hue): Topic keywords and temporal trends of the "Information retrieval and data mining" research domain. Bottom (with a green hue): Topic keywords and temporal patterns of the "Human Computer Interaction" field.

The second column in Figure 9 illustrates the overall temporal evolution of selected groups. The user noticed that the trend of proposals awarded under the *IR* group seemed steady with a slight decline over the recent two years. To examine and compare the development of individual topics in the IR group, the users further isolated three topics that are of interest. The corresponding trends for these topics are shown next to the overall trend.

Through quickly examining the volume of each topic trend, the user confirmed his hypothesis that topic 18 on "web search and document retrieval" has continued to be a more popular subfield over the years in terms of NSF research funding (Figure 9 ribbon with red border). However, the user was also surprised when found out that the *"HCI"* group exhibits a slight decline in recent years after a steady growth around 2007. Through examining individual topic trends, more interesting patterns prevailed. Although the overall trends for other topics group have subsided slightly, the research on"*affective computing and emotion related studies*" has gone up significantly in the past two years, as outlined in Orange.

This use case illustrated that the visual interface not only enables the user to view trends for a group of topics that describe a research field, but also permits the discovery of the contributions of individual topics to the overall trends as well as anomalies. According to the user, such analysis gave him valuable insights in understanding the research trends in the areas he is interested in and could potentially help him adjust future proposal focus.

4.2 Identifying program impacts in research

Given that the above two topic groups all exhibits slight downward trending, the user wanted to identify upcoming research topics that received more funding interest in the recent years. He started by mouse hovering over each topic ribbon in the main Hierarchical ThemeRiver view, looking for increasing trends.

Two topic groups caught his attention as shown in Figure 8. Both groups exhibit increasing volume in the past three years, indicating more research proposals were awarded in the two areas. The top row illustrates a topic group related to environmental related research as well as citizen science. As shown in the individual temporal trend for each topic, the user identified that the topic on citizen science and spatial temporal analysis significantly contributed to the recent growth of the focused topic group.

The second row in Figure 8 illustrates a topic group that summarizes research on medical and healthcare related research. Through enabling the time column selection, the user selected proposals related to the health care topic that were awarded in 2012, highlighted in the yellow rectangle. He then discovered that most of the proposals were related to health monitoring and were awarded by the only-recently launched program–Smart and Connected Health (2011).

The user was pleased to find out the impact of a newly established program on research trends and considered the HierachicalTopics a powerful tool in aiding the discovery of the contributors to the temporal changes and possibly the cause for such changes.

5 User Study

To quantitatively evaluate the utility of HierarchicalTopics in aiding users analysis of a text corpus, we conducted a formal user study focusing on comparing hierarchical to non-hierarchical topic structure. Our hypothesis is that the hierarchical topic structure would yield faster identification of topics that are similar in nature.

5.1 Data and Tasks

The dataset used for the user study contains 2453 news articles published between Sept 2012 to March 2013 on CNN.com. Two conditions were designed to evaluate the effect of hierarchical topic structure versus representing them as a flat list of topics. We designed two tasks for the experiment: the first task aims to group individual topics into different news categories; the second task focuses on examining the overall temporal trends for the topics in each news category. For the second task, we required the participants to group all the topics based on their findings in task 1.

Specifically, in task 1, we asked the participants to identify news topics that fall into the following five categories: American Politics, Sports and Entertainment, Natural Disaster, Health-Related Issues, and Middle-East News. An example topic grouping result produced during one of the experiments is shown in Figure 1. Each participant was provided an answer sheet to write down the topic number belonging to each category. For each topic the participants have identified, we also asked them to provide a score (1-5, with 5 as very confident) indicating their confidence of how much the topics fits into their category of choice. For the second task, we asked the participants to group the topics identified in task 1 based on their category. The grouping was done through drag-and-drop interactions within the visual interface. After each group of topic has established, we asked the participants to examine and describe the temporal trend for the topic groups.

To control the complexity of the tasks, we extracted 40 topics from the news corpus. The reason for doing so was that the participants assigned to the non-hierarchical topic organization had to go through the topics one by one. With no initial aid of organizing similar topics together, grouping large number of topics would become laborious and require a lot of repetitions of the same operations. This implies that, if the hierarchical structure proves superior in this study, it will increase its edge relative to a flat structure as the number of topics grows.

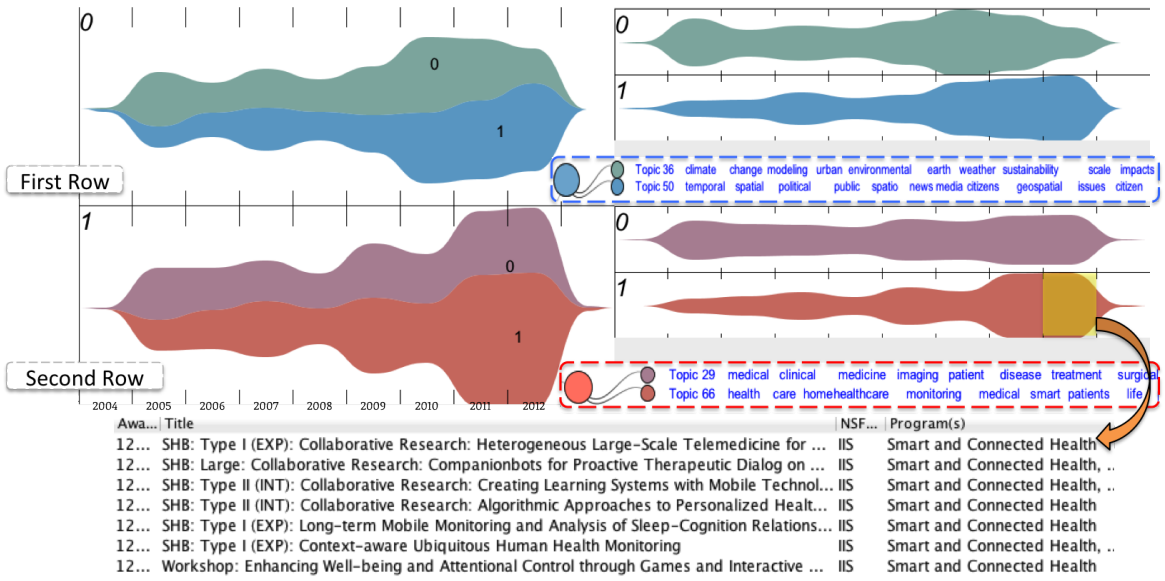

Fig. 8. Case Study: Making sense of increasing topic group trends. Top (with a blue hue): topic group of "environmental and citizen science" has seen recent growth. Middle (with a red hue): heath care related topic group exhibit growth in the past two years, with the "health monitoring" topic as the major contributor to the overall growth. Bottom: detail view showing proposals regarding the "health monitoring" topic awarded in 2012.

5.2 Experiment Design

Eighteen participants took part in the study (13 male, 5 female). The age of the participants ranged from 18 to 34.The study used a between-subjects design. All participants were first provided 10 minutes of training on the HierarchicalTopics visual interface. Each participant was then randomly assigned to one of two conditions (hierarchical vs. non-hierarchical topic organization). The participants were asked to write down their findings on an answer sheet, which records the identified topic numbers for each listed category for the first task and the pattern of the temporal trends for the second task. The experimenter timed the participants for completing each category while they were performing the tasks. The study was conducted in a lab setting, on a computer with two displays (resolution at 2560x1600 and 1920X1200, respectively), 2x 2.66GHz CPU and 12 GB memory.

5.3 Results

For the purpose of analyzing whether the hierarchical topic structure helps the analysis of large text corpora, we calculated the difference of average time for identifying topics for each news category. The average time is computed as the overall time to find all topics for each category, divided by the number of topics identified. The reason for using the average time is because participants identified different number of topics for a given category. In practice, determining whether a topic belongs to a certain category can be subjective. For instance, some participants consider a topic related to the trial of Conrad Murray (the physician for Michael Jackson) belonging to the "Sports and Entertainment" category since it's related to the pop singer. Other participants may consider this being a stretch since Michael Jackson is not the main subject of the news articles related to the topic.

For the same reason, we did not grade the accuracy of the identified topics, since arguments could be made for topics to be included or excluded from a news category. Although we did not grade accuracy of the identified topics, most of the identified topics for each news category did overlap. Two experimenters independently examined each participants' answer, and they did not find answers that are clearly not pertinent to the categories.

5.3.1 Speed: hierarchical topic vs. non-hierarchical topic organization

To measure whether the hierarchical topic organizations yield faster speed for identifying topics for each news category, we performed one-way ANOVA on each category. A significant effect was found for two categories: American Politics and Middle-East News. For the American Politics category, a significant effect of hierarchical topic organization on the time for identifying relevant topics (Task 1) was found at the $p<.05$ level for the two conditions [$F(1,16) = 4.84$, $p = .043$]. For the same category, a significant effect was also found between two conditions [$F(1,16) = 4.79$, $p = .044$] in task 2, which involves grouping the identified topics and observing the temporal trends. For the Middle-East News category, the ANOVA revealed a significance between two conditions [$F(1,16) = 5.15$, $p= .37$]. No significance was found for the other three categories. Detailed results are shown in Figure 10 (left).

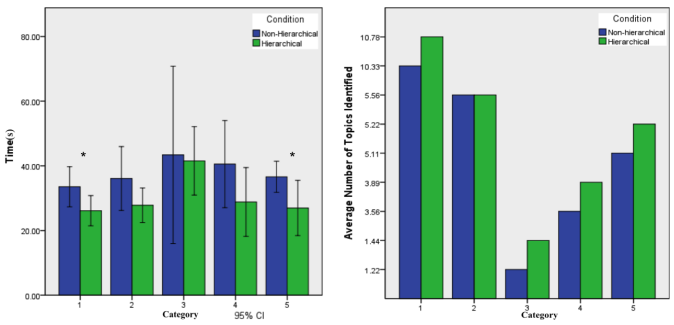

Fig. 10. Left: Average time to identify all topics for each news category during task1. Asterisk denotes significant difference. Right: Average number of topics identified for each news category.

Combining with the average number of topics found in each category shown in Figure 10 (right), the results became more informative. Significant differences were found for categories with relatively large number of topics. In other words, the hierarchical topic structure lead to faster identification and grouping of large number of relevant topics.

5.3.2 User's confidence and Response on potential scalability of the system

As mentioned in section 5.1, during task 1, when a participant jotted down the topics for each category, we have also asked her to provide a confidence value of how well the topic fits into the category. The confidence values for all participants assigned to the hierarchical condition

have a mean of 4.5, with a standard deviation of 0.52. The confidence values for participants assigned to the other condition exhibit a mean of 4.47, and a standard deviation of 0.5. Although no statistical significance was found, the participants under the hierarchal condition consistently reported higher average confidence value for each news category. Note that with 5 as the most confident, the mean values of the confidence show that all participants are fairly certain about their answer. From another perspective, the high confidence values also reflect that the participants could interpret the topics and possibly the topic hierarchies without much difficulty.

The last question on the answer sheeting was regarding the potential scalability of the system. In particular, the question asked the participants to comment on if the HierarchicalTopics could scale to hundreds of topics. We tallied the participants' response. 4 out of 9 participants assigned to the hierarchical condition answered "yes", 4 answered "maybe", and the rest 1 participants answered "no". In contrast, 0 out of 9 participants assigned to the non-hierarchical condition answered "yes" to potential scalability, while 6 answered "maybe" and 3 answered "no".

None of the participants assigned to the non-hierarchical topic condition thought the system could scale to hundreds of topics, while the participants answering "*Maybe*" under the same condition further commented that some sort of automated classification such as topic groups could make the system much more scalable. The participants assigned to the hierarchical topic condition provided more positive responses toward the potential scalability of our system. Several of constructive comments were generated based on user feedback, details of which will be described in the discussion session.

In summary, the study results reveal that hierarchical topic structure leads to more efficient identification and grouping of larger numbers of relevant topics. After performing two tasks through interacting with the visual interface, most participants consider the hierarchical system scalable and bears potential to handle hundreds of topics.

6 Discussion

In this section, we discuss possible improvements on the topic rose tree algorithm and the visual interface.

6.1 Implicit modeling assumptions and design elements

One implicit assumption of organizing large number of topics into a hierarchy is that the topics can fit cleanly into such a structure. However, in practice, such assumption may not always hold. For example, certain topics may fit into multiple groups based on users' interpretation. To address this issue, we could allow users to duplicate topics and add the topics into the corresponding groups.

Another implicit assumption is that we assume that the topic results are fine-grained enough so that the "split" operation is currently not supported in the HierarchicalTopics system. We think the "split" operation is potentially very important since it permit users to directly influence the topic models. However, there are several reasons that the splitting operation is challenging to support. First, asking users to specify how to split the topics (words that should or should not be group) could quickly turn into a laborious task if the interactions are not properly designed. Second, since the the computation of topics usually involves hundreds of interactions, rebuilding the topic model based on users' input of how to split the topics is difficult to achieve in real time [17]. Despite the challenges, we consider the "split" operation a very important option, and a great contribution for interactive visualization to potentially bring to the topic modeling community. Therefore, our future work will try to address this issue and more broadly to permit users to modify the underlying topic model in real or semi-real time.

6.2 Limitation and future improvements on HierachicalTopics system

During the study, the participants provided constructive comments for improving HierarchicialTopics. A few users mentioned the need for annotation feature, which would allow them to annotate or bookmark a general topic group. In addition, users would also like to search for a particular word in the topic view, for the purpose of discovering all topics containing a word of interest. As mentioned in Section 3.3.1, we have already incorporated both the annotation feature and the search function into the current system based on the feedback.

Another interesting comment was on possibly taking advantage of spatial organization of the topics. One participant would like to organize the topics into interested vs. not interested piles and place them on different parts of the screen. Spatial organization is commonly used when working with real objects, and has been shown to aid more complex sense-making processes [1]. Thus more flexible user interactions need to be supported for users to accomplish such task in an un-laborious manner.

During the study, a few participants raised the question of what if one topic falls into two or more topic groups. For example, the topic of human robot interaction could be categorized into both HCI related topic group and Robotics related group. Therefore, we are planning to provide additional user interactions that allow users to duplicate topics and keep track of the duplicates.

Lastly, one limitation arose from the use of tree visualization to represent the hierarchical topic structure. The concern is that tree visualizations may not scale to displaying very large number of topics or multi-level hierarchies. Our HierarchicalTopics system alleviates this issue by supporting multiple user interactions, including collapsing, annotating, and deleting the nodes in the rose tree. Nonetheless, we acknowledge the potential limits of this tree representation and will further explore other visual metaphors.

6.3 Future improvement on the Topic Rose Tree

As of the Topic Rose Tree algorithm, improvements could be added to make the algorithm more transparent and interactive to end-users. For example, when merging two subtrees in each computational step, selecting different operations would yield different results not only in terms of topic groups, but also regarding the depth of the tree. Theoretically, both the absorb and collapse operations would lead to a rose tree with smaller depth compared to the join operation. Trees with less depth may make more sense for grouping topics, since the topics were assumed to be equally descriptive in the topic models. In the hLDA [3], topics on a higher level are usually less meaningful, comprised of mainly stopwords. Thus it makes sense to control the tree depth to be as small as possible. A simple way to influence the depth of Topic Rose Tree is to encourage the absorb and collapse operation rather than the join operation. New interactions could, therefore, be designed to allow users to tweak the weight when calculating the cost of each operation. Such interactions could potentially support advanced users in influencing the topic hierarchy generation. This will be one of the future directions for our visual text analytics research.

7 Conclusion

In this paper, we present HierarchicalTopics, a visual analytics approach to support the analysis of text corpora based on large number of topics. HT is designed to address three challenges faced when analyzing large text corpora through topic based methods. HierarchicalTopics not only provides initial hierarchical structure of topics to facilitate exploration and navigation, it further allows users to modify topic hierarchies based on users' interest through intuitive interactions. In addition, the ThemeRiver in HierarchicalTopics is tailored to represent temporal trends in a hierarchical fashion. It enables the analysis and comparison of groups of topics as opposed to viewing the evolution of one topic at a time. Through both case study and user experiments, we have demonstrated the efficacy of HierarchicalTopics in helping users identifying topics groups, as well as interesting temporal patterns.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Science Foundation under award number SBE-0915528 and DHS VACCINE Center of Excellence.

REFERENCES

[1] C. Andrews, A. Endert, and C. North. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 55–64, New York, NY, USA, 2010. ACM.

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

[3] D. M. Blei, T. Gri, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *Neural Information Processing Systems(NIPS)*, 2003.

[4] D. M. Blei and J. D. Lafferty. Correlated topic models. *Neural Information Processing Systems*, 2006.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[6] C. Blundell, Y. W. Teh, and K. A. Heller. Discovering nonbinary hierarchical structures with bayesian rose trees. *Mixtures: Estimation and Applications*, April 2011.

[7] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In *Neural Information Processing Systems (NIPS)*, 2009.

[8] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *IEEE VAST*, pages 143–152, 2012.

[9] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces*, 2012.

[10] J. Chuang, D. Ramage, C. D. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *ACM Human Factors in Computing Systems (CHI)*, 2012.

[11] CNN. Library of congress digs into 170 billion tweets. http://bit.ly/Uwqi7X.

[12] Committee on National Statistics. Science of science and innovation policy principal investigators' workshop. http://bit.ly/10o3via, Sep 2012.

[13] 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. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2412–2421, 2011.

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

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

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

[17] Y. Hu, J. Boyd-Graber, and B. Satinoff. Interactive topic modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 248–257, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[18] A. Jinha. Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010.

[19] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. ivisclustering: An interactive visual document clustering via topic modeling. *Comp. Graph. Forum*, 31(3pt3):1155–1164, June 2012.

[20] Medialab Tools. i want hue web color chooser. http://tools.medialab.sciences-po.fr/iwanthue/, March 2013.

[21] J. Paisley, C. Wang, and D. M. Blei. The discrete infinite logistic normal distribution for mixed-membership modeling. *Bayesian Analysis*, 7(4):997–1034, 2012.

[22] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. AAAI, 2010.

[23] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 457–465, New York, NY, USA, 2011. ACM.

[24] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[25] D. Shahaf, C. Guestrin, and E. Horvitz. Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 1122–1130, New York, NY, USA, 2012. ACM.

[26] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. Zhou. Understanding text corpora with multiple facets. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 99–106, 2010.

[27] R. M. Shiffrin and K. Börner. Mapping knowledge domains. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5183–5185, 2004.

[28] Statisticbrain.com. Facebook statistics. http://bit.ly/YaAVmg.

[29] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004.

[30] The National Science Board. *Science and Engineering Indicators 2010, Chapter 5, Page 29*. National Science Foundation, 2010.

[31] The Unofficial Twitter Resource. Twitter now seeing 400 million tweets per day, increased mobile ad revenue, says ceo. http://bit.ly/JP9DXA, Feb 2013.

[32] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 448–456, New York, NY, USA, 2011. ACM.

[33] X. Wang, W. Dou, Z. Ma, J. Villalobos, Y. Chen, T. Kraft, and W. Ribarsky. I-SI: Scalable Architecture of Analyzing Latent Topical-Level Information From Social Media Data. *Computer Graphics Forum*, 31(3):1275–1284, 2012.

[34] 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 *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 153–162, New York, NY, USA, 2010. ACM.

[35] ZD Net. Engaging citizens the right way: Government uses twitter during hurricane irene. http://zd.net/mS0aOU, Sep 2011.