ParallelTopics: A Probabilistic Approach to Exploring Document Collections

Wenwen Dou* UNC Charlotte
Xiaoyu Wang† UNC Charlotte
Remco Chang‡ Tufts University
William Ribarsky§ UNC Charlotte

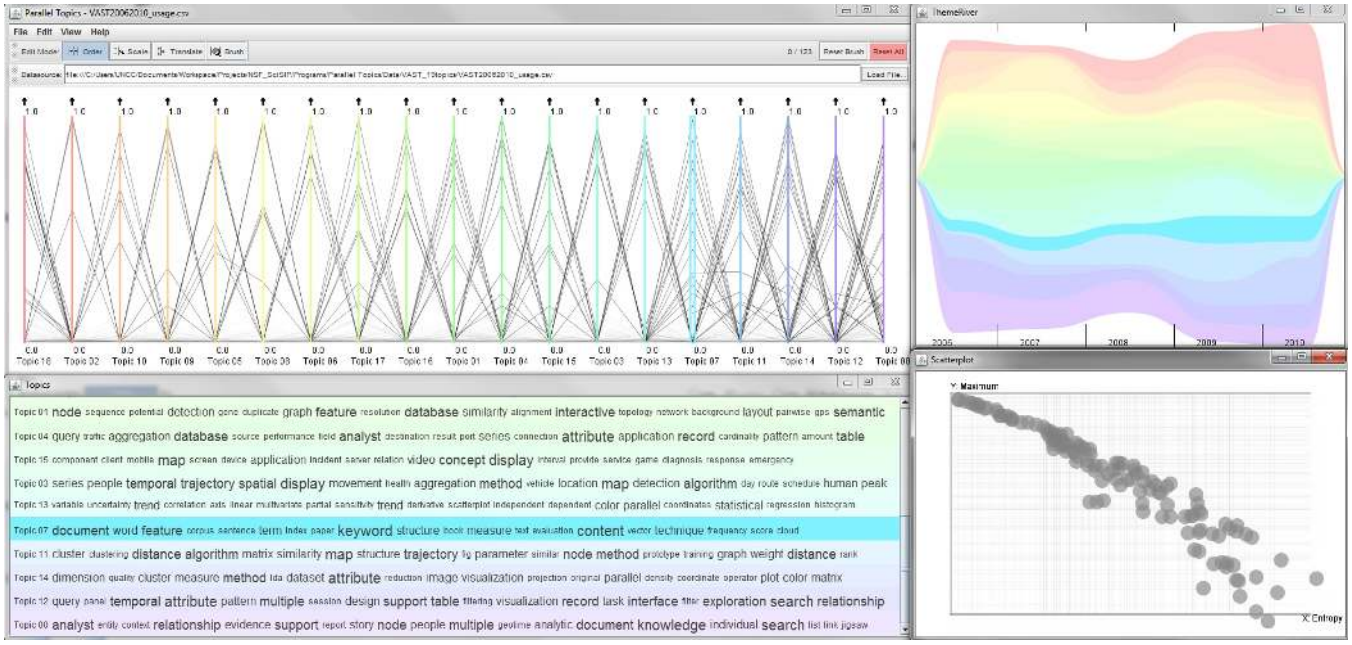

Figure 1: Overview of ParallelTopics. Topic07 is highlighted. Top left: Document Distribution view, top right: Temporal view, bottom left: Topic Cloud, bottom right: Document Scatterplot. A user is hovering the mouse over Topic 07 (light blue) in the Document Distribution view.

ABSTRACT

Scalable and effective analysis of large text corpora remains a challenging problem as our ability to collect textual data continues to increase at an exponential rate. To help users make sense of large text corpora, we present a novel visual analytics system, ParallelTopics, which integrates a state-of-the-art probabilistic topic model Latent Dirichlet Allocation (LDA) with interactive visualization. To describe a corpus of documents, ParallelTopics first extracts a set of semantically meaningful topics using LDA. Unlike most traditional clustering techniques in which a document is assigned to a specific cluster, the LDA model accounts for different topical aspects of each individual document. This permits effective full text analysis of larger documents that may contain multiple topics.To highlight this property of the model, ParallelTopics utilizes the parallel coordinate metaphor to present the probabilistic distribution of a document across topics. Such representation allows the users to discover single-topic vs. multi-topic documents and the relative importance of each topic to a document of interest. In addition, since most text corpora are inherently temporal, ParallelTopics also depicts the topic evolution over time. We have applied ParallelTopics to exploring and analyzing several text corpora, including the scientific proposals awarded by the National Science Foundation and the publications in the VAST community over the years. To demonstrate the efficacy of ParallelTopics, we conducted several expert evaluations, the results of which are reported in this paper.

Index Terms: H.5.2 [INFORMATION INTERFACES AND PRESENTATION]: User Interfaces—Graphical user interfaces (GUI);

1 INTRODUCTION

The management of large and growing collections of text information is a challenging problem. Data repositories of knowledge-rich texts have become widely accessible, leading to an overwhelming amount of information to organize and explore. As the number of documents increases, identifying the gist of the corpora becomes cognitively costly and time consuming.

The challenge of automated summarization of large text corpora has been a primary area of interest for researchers in the natural language processing (NLP) domain. To summarize a text corpus, researchers have developed techniques such as Latent Semantic Analysis (LSA) for extracting and representing the contextual-usage meaning of words [21]. The LSA produces a concept space which could then be used for document classification and clustering. More recently, probabilistic topic models have emerged as a powerful new technique for finding semantically meaningful topics in an unstructured text collection [6]. To further provide a visual

*e-mail: wdou1@uncc.edu
†e-mail:viztang@gmail.com
‡e-mail:remco@cs.tufts.edu
§e-mail: ribarsky@uncc.edu

summary of text corpora, researchers from the knowledge discovery and visualization community have developed tools and techniques to support visualization and exploration of large text corpora based on both LSA (e.g. [32, 12]) and topic models (e.g. [19, 22, 30, 24]).

Although probabilistic topic models have demonstrated their advantages in interpretability and semantic association [15], few interactive visualization systems have leveraged such models to support exploration and analysis of text corpora. The exemplar-based visualization [11] and probabilistic latent semantic visualization [19] projected documents onto static 2D plots while estimating topics of a text corpus. Although the clusters of documents conform well to the labels, there is little room for interactive exploration and analysis of the document clusters. One exception is the time-based visualization system TIARA [22, 30, 29], which applies the ThemeRiver [16] metaphor to visually summarize a text collection based on the topic content. Through analysis with the TIARA system, users could answer questions such as: What are the major topics in the document corpus? and How have the topics evolved over time?

However, when analyzing large text corpora, there are many other real-world questions that current text analysis visualization systems have difficulty answering. In particular, questions pertaining to the relationships between topics and documents are difficult to answer with existing tools. Such questions include: what are the characteristics of the documents based on their topical distribution? and what documents contain multiple topics at once (and what are they)? In the field of science policies, documents with multiple topics could indicate publications that are interdisciplinary (i.e. that cover more than one body of knowledge). Similarly, in the context of social media analysis, a document with multiple topics may signify a unique news article that is relevant to different hot topics.

To address such needs in real-world applications, we have developed a visual analytics system ParallelTopics, which tightly integrates interactive visualization with a state-of-the-art probabilistic topic model. Specifically, in order to answer previous questions, ParalleTopics utilizes the Parallel Coordinate metaphor to present the probabilistic distribution of a document across topics. This carefully chosen representation not only shows how many topics a document is related to, but also the importance of each topic to the document of interest. Moreover, ParallelTopics provides a rich set of interactions that can help users to automatically divide a document collection based on the number of topics in the documents. In addition to depicting the relationships between topics and documents, ParallelTopics also supports other tasks, which are also essential to understanding a document collection, such as summarizing the document collection into major topics, and presenting how the topics evolve over time.

To summarize, the set of questions that ParallelTopics can effectively address when analyzing large text corpora include:

- Q1: What are the major topics that well capture the document collection?
- Q2: What are the characteristics of the documents based on their topical distribution?
- Q3: What documents address multiple topics at once?
- Q4: How do the topics of interest evolve over time?

To help users answer these questions, ParallelTopics first extracts a set of semantically meaningful topics using the Latent Dirichlet Allocation (LDA) mode [9]. To support visual exploration of a document collection based on the topic model, ParallelTopics employs multiple coordinated views to highlight both topical and temporal features of document corpora. The novel contribution of ParallelTopics lies in the depiction of the probabilistic distributions of documents over topics and supporting interactive identification and more detailed examination of single-topic and multi-topic documents. To evaluate the efficacy of the ParallelTopics, we conducted an evaluation with several expert users on two different text corpora. Our evaluation indicates that ParallelTopic is effective in addressing the four intended questions in the context of specific domains.

2 Related Work and Background

Two lines of work, namely text analysis models and text visualization techniques, were the main inspiration for the design of ParallelTopics.

2.1 Text processing

The first major progress in text processing was due to the vector space model [28]. In this model, a document is represented as a vector in a high-dimensional space where each dimension is associated with one unique term within the documents. One well-known example of VSM is the tf-idf [27], which evaluates how important a word is to a document in a corpus. Although the VSM has empirically shown its effectiveness, it suffers from a number of inherent shortcomings to capture inter- and intra-document statistical structure [3].

To address the shortages of the VSM, researchers have introduced latent semantic analysis (LSA) [21], which is a factor analysis that reduces the term-document matrix to a much lower dimension subspace that captures most of the variance in the corpus. Although LSA overcomes some of the drawbacks from the VSM, it has its limitations as well [3]. The new feature space is difficult to interpret since each dimension is a linear combination of a set of words from the original space.

Being aware of the limitations of LSA, researchers have proposed generative probabilistic models to document modeling. Blei et al. introduced a generative model that represents the content of words and documents with probabilistic topics instead of a purely spatial representation [9]. One distinct advantage of such representation is that each topic is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms [5]. The LDA model postulates a latent structure consisting of a set of topics; each document is produced by choosing distribution over topics, and then generating each word at random from a topic chosen by using this distribution. The extracted topics capture meaningful structure in the otherwise unstructured data, as shown in analyzing scientific abstracts [14] and newspaper archives [31]. On a cognitive level, the LDA model performs well in predicting word association and the effects of semantic association and ambiguity on a variety of language-processing and memory tasks [15].

Because of various advantages of the LDA model, ParallelTopics first utilizes the model to extract a set of semantically meaningful topics given a text corpus. ParallelTopics then present the probabilistic results in an intuitive manner to make the complex model easily consumable by users when analyzing large text corpora.

2.2 Visualization of text corpora

Despite advances in automatic text processing techniques, human intelligence still plays a key role when analyzing text corpora. Therefore, a number of visualization systems and techniques have been developed based on the text processing methods to keep users in the loop.

Utilizing the VSM model, Themail was introduced by Viegas et al. to visualize email content with the purpose of portraying relationship from conversational histories. The keywords within the visualization are generated based on the tf-idf algorithm.

Storylines [32] enables users to visually explore text corpora through a social network metaphor based on latent semantic analysis results. Other visualization systems have used multidimensional

projection methods (e.g. PCA [25], MDS [20]) to visualize text corpora. These projection techniques are similar to LSA in spirit since they represent the documents as vectors with term frequency as their features and then identify a lower-dimension projection space [11]. Visualization systems based on these projection techniques include IN-SPIRE [1] and Infosky [4]. More recently, to visualize large classified document collections, Oesterling et al. [24] proposed a two-stage framework for a topology-based projection and visualization. Unlike most traditional clustering techniques in which a document is assigned to a specific cluster, ParallelTopics accounts for different topical aspects of each individual document.

Since the debut of topic models, visualization systems have utilized such models for their advantages over previous text processing techniques. The exemplar-based visualization [11] and probabilistic latent semantic visualization [19] projected documents onto static 2D plots while estimating topics of a text corpus. Although the visual clustering results are better than the ones obtained from the multidimensional projection methods, there are several limitations. First, as the number of extracted topics grows, the document clusters in the 2D projection are no longer separable based on topics. What is more, there is little room in these visualization tools for interactive exploration and analysis of the document clusters. Most recently, Wei et al introduced TIARA, a time-based interactive visualization system that presents the extracted topics from a given text corpus in a time-sensitive manner [30]. TIARA provides a good overview of the topics with respect to their evolution over time. However, the relationship between documents and topics is less clear.

In ParallelTopics, we present the probabilistic distribution of documents across the extracted topics in addition to describing topic evolution over time. Thus we provide an overview of the characteristics of documents based on their topical distribution and enable users to identify documents that address multiple topics at once.

3 PARALLELTOPICS

ParallelTopics supports the exploration of a document collection on multiple levels. On the overview level, the system assists users in answering questions such as: What are the major topics of the document collection? (Q1) and What are the characteristics of the documents in this collection? (Q2) On the facet level, ParallelTopics supports activities such as identifying temporal trend (Q4) of a specific topic and identifying documents that are related to multiple topics of interest (Q3). On the detailed level, the ParalllelTopics system allows access to details of each individual document on demand.

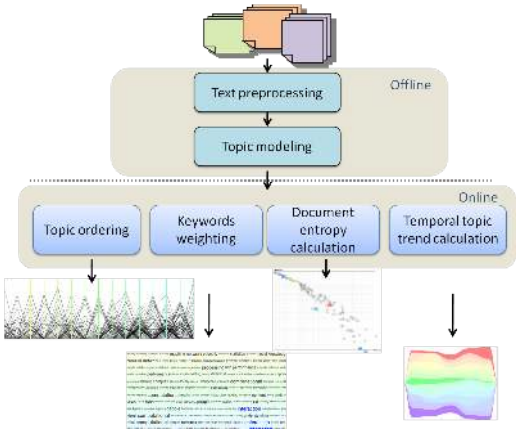

Figure 2: System architecture of ParallelTopics.

Based on the state-of-the-art topic model, ParallelTopics employs multiple coordinated views with each view addressing one of the aforementioned questions. In this section, we describe the design of the ParallelTopics system. Figure 2 illustrates the overall architecture of ParallelTopics. Starting from the top, document processing and topic modeling are done offline. Based on the offline processing results, each online module serves one specific visualization view in the ParallelTopics. We start by introducing the topic model that underpins the ParallelTopics system.

3.1 Topic-based text summary

As described in section 2.1, topic models have several advantages over traditional text processing techniques. Therefore we employ a probabilistic topic model in the ParallelTopics to summarize document collections. More specifically, we used Latent Dirichet Allocation (LDA) [9] to first extract a set of semantically meaningful topics. LDA generates a set of latent topics, with each topic as a multinomial distribution over keywords, and assumes each document can be described as a probabilistic mixture of these topics [10]. To introduce the notation, we write $P(z)$ for the distribution over topics z in a particular document. We assume that the text collection consists of D documents and T topics. These notations will be used throughout the rest of paper.

In our system, we first process the document collection and remove stopwords such as in IN-SPIRE [1]. We then use the Stanford Topic Modeling Toolbox (TMT) [26] to extract a set of topics from the document collection. The extracted topics and probabilistic document distributions serve as input to the visualizations in the ParallelTopics.

3.2 Interactive visual exploration of text corpora

In this section, we introduce the visual design of ParallelTopics. The system consists of four coordinated overviews: a Document Distribution view that displays the probabilistic distribution of documents across topics; a Topic Cloud that presents the content of the extracted topics; a Temporal view that highlights the temporal evolution of topics; and a Document Scatterplot that facilitates interactive selection of single-topic vs. multi-topic documents. Each of the four views in the ParallelTopics system serves a distinct purpose, and they are coordinated through a rich set of user interactions. In addition, upon selection of any documents, we provide a Detail view that presents the text content on demand.

3.2.1 Topic Cloud : Revealing the major topics (Q1)

To help the users quickly grasp the gist of a document collection, we present the topics as a tagcloud. In the Topic Cloud view, each line displays a topic, which consists of multiple keywords. Since each topic is modeled as a multinomial distribution over keywords, the weight of each keyword indicates its importance on the topic. To encapsulate such information in the Topic Cloud, we align the keywords from left to right with the most important keyword at the beginning. In addition, since one keyword may appear in multiple topics, the size of each keyword reflects its occurrences within all topics. An example of the Topic Cloud view is shown in figure 3. To assist users in understanding the major topics in a document collection, we present the topics in a sequence that semantically similar topics are close by so that there is continuity when scanning over the topics sequentially. Since the LDA model does not model the relationship between topics, we reorder the topics by defining a similarity metric. The details of the reordering is described in section 3.2.2.

The Topic Cloud view also provides users with a set of interactions to help users quickly make sense of the topics. For example, hovering over a particular keyword would highlight all other occurrences in the Topic Cloud. A user may also search for a particular keyword of interest. In addition, the Topic Cloud view is tightly

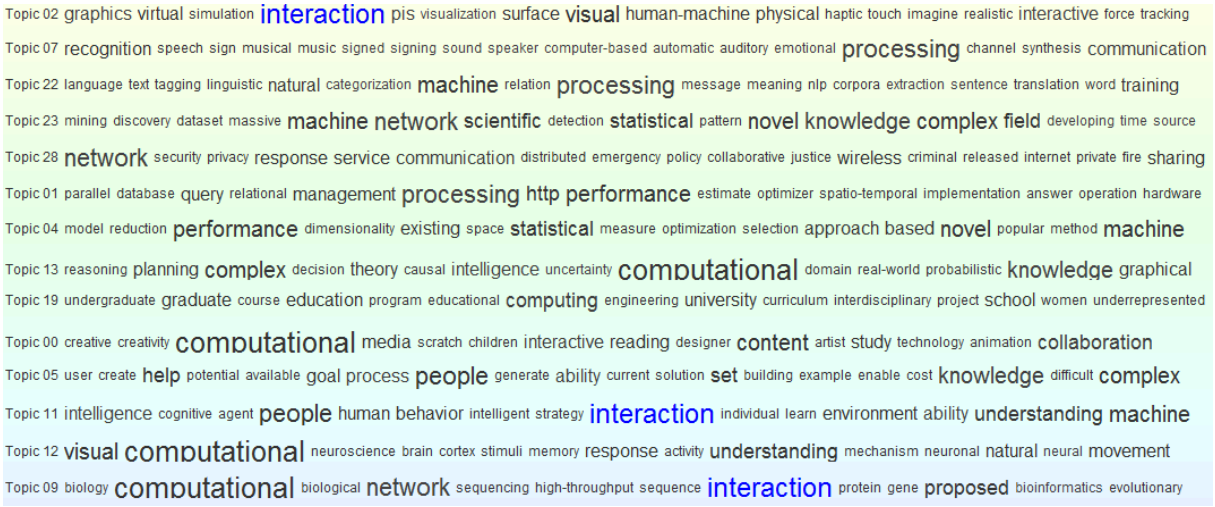

Figure 3: Topic Cloud with keyword "interaction" highlighted. Each "interaction" has a slightly different meaning under the context of its topic. The topics are extracted from the abstracts of proposals awarded by NSF from 2000 to 2009.

coordinated with all other views to promptly provide information regarding a specific topic on demand.

3.2.2 Document distribution: Depicting the characteristics of the documents (Q2)

To provide an overview of documents as mixtures of topics, we highlight the distribution of each document across all extracted topics. Our representation converts the documents probabilistic distributions to signal-like patterns that signify each document. More specifically, we adopt the parallel coordinate metaphor [18] with each axis denoting a topic and each line representing a document in the collection (figure 1, top left view). In our use of the parallel coordinate, all variables (topics) are uniformly spaced, and every variable share the same value range from 0 to 1. Therefore, when viewing the document distribution view, it is not necessary to make sense of a document based on its value on each individual axis but based on the pattern across all the axes as a whole. In the following subsections, we first introduce an important consideration regarding the visual order of the axes. We then categorize the characteristics different documents may present across topics.

Topic ordering One limitation of LDA is that it does not directly model the correlation between topics, but in most text corpora, it is natural to expect the correlation between the occurrences of topics [6]. We want to address this limitation so that similar topics appear next to each other in the visualization. Coincidentally, one characteristic of the parallel coordinate visualization is that correlation between adjacent axes are much easier to discover [13]. Therefore we use Hellinger distance to order the topics so that the correlation between similar topics become visually salient. We consider the distribution of each topic across all documents as a probability distribution $f(x)$. We then define topic similarity as the Hellinger distance between two topic distribution among all documents:

$$distance(i,j) = \frac{1}{2}\int (\sqrt{f_i(x)} - \sqrt{f_j(x)})^2\, dx \qquad (1)$$

Here $f_i(x)$ and $f_j(x)$ are the ith and jth topic probabilistic distribution over all D document in the entire collection. Therefore, the derived distance measures how similar any two topics are given a text corpus. When plotting the topics as axes in our interface, we start with a topic with the most probabilistic concentration and then always look for the most similar topic to the current one based on their distances. Figure 4 demonstrates the visualization of documents across topics after topic reordering. The relationship between any two most similar topics (adjacent axes) becomes visually identifiable.

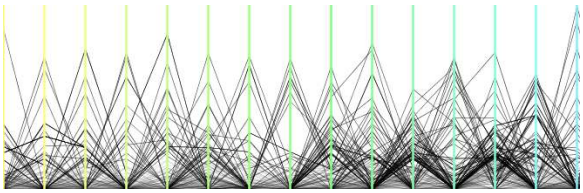

Figure 4: Similar topics are ordered to be next to each other so that the relationship between them are visually identifiable.

Document Characteristics When exploring the document distribution over topics, one can easily discover that documents present different characteristics based on the number of topics they have. Figure 5 illustrates documents that focus on only one topic, two topics, and more than two topics. Different number of topics within documents can be interpreted as distinct characteristics given a context of the text collection. For example, in a collection of scientific publications, documents with one topic denote publications on a specific research field. Documents with two or more topics are more likely to represent interdisciplinary research articles, which often integrate two or more bodies of specialized knowledge [2].

In addition, the document distribution view provides a rich set of interactions, such as brushing, highlighting, etc. Brushing a probability range on a topic allows users to select documents that have a certain probability for that specific topic. Through synthesizing the information from both Topic View and Document Distribution View on the major topics and document characteristics, a user could effectively develop an overview of the document collection.

3.2.3 Document Scatterplot: Investigating documents based on their number of topics (Q3)

The document distribution view enables users to identify documents that focus on a specific topic through brushing the top range on the topic. However, identifying documents that are related to two or more topics in a large corpus is not as straightforward since

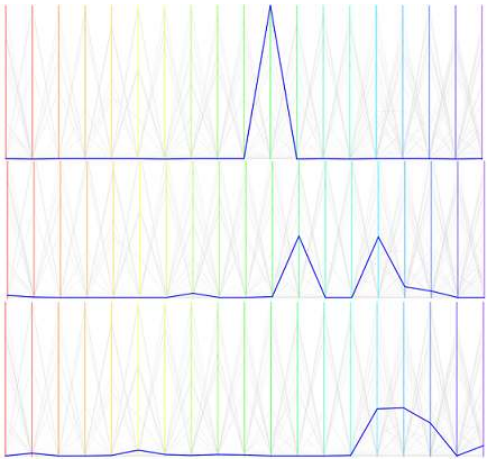

Figure 5: Document Characteristics: Top - Single Topic document; Middle - Bi-topic document; Bottom - Multi-topic document.

they are shadowed by high probability values of the single-topic documents. To alleviate this problem, we represent all documents in a way that single-topic and multi-topic documents are easily separable.

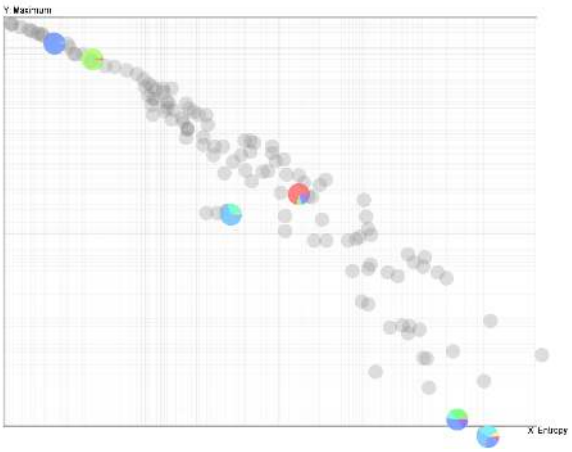

Figure 6: Document Scatterplot: the position of each document in the scatterplot correlates to its number of topics. Single-topic documents are in the top left corner while multi-topic documents reside in the bottom right corner. Each pie glyph is colored based on number of topics in each document.

Document Entropy As seen in the Document Distribution view, each document is converted into a signal-like probabilistic distribution pattern (figure 1). In this representation, documents with multiple topics appear noisier than the ones that clearly focus on one topic. In information theory, Shannon entropy is a measure of the amount of uncertainty associated with a random variable. Assuming the topic as a random variable for each document in our context, Shannon entropy could be used to separate clear signal from noisy ones. Therefore, we applied Shannon entropy to distinguish documents based on the number of topics they have. We calculate the entropy of each document based on their probabilistic distribution across topics:

$$H_k = -\sum_{i=1}^{T} P(d_k|z=i) log_2 P(d_k|z=i) \quad (2)$$

Here, $P(d_k)$ is the probabilistic distribution of the kth document over all topics. We then plot each document based on its entropy and its maximum probability value over topics (normalized to [0, 1]) in a scatterplot view. In this presentation, single-topic documents (with higher max value and lower entropy) are at the top left corner within the scatterplot while bottom right corner captures documents with more number of topics (lower max value and higher entropy). Upon selection, pie glyphs are shown to describe the topical contribution to a specific document. In figure 6, each pie glyph represents a selected document, with each color denoting a topic. As shown, documents with smaller entropy values appear as pie glyphs as a solid circle; whereas documents with larger entropy values appear to have multiple colors, indicating that entropy values do correspond to the number of topics in the input documents.

In summary, the Document Scatterplot enables users to interactively identify subgroups of documents with desired number topics through selecting document within different regions.

3.2.4 Temporal View: Presenting topic evolution over time (Q4)

Since most document collections are accumulated over time, it is helpful to present such temporal information to assist users in understanding how topics of a corpus evolve. Our Temporal view is created as an interactive ThemeRiver [16], with each ribbon denoting a topic (figure 7). In the text corpus, each document is associated with a time stamp, therefore the height of each ribbon over time could be determined by summing document distribution on this topic within every time frame. The unit of time frame depends on the corpora, for example, year might be a proper time unit for scientific publications while month or even date would be more appropriate for news corpora. After the time unit has been chosen, we divide the documents into the corresponding time frame based on the time stamp. Then for each time frame, we calculate the height of each topic by summing the distribution on the topic from the documents within the time frame.

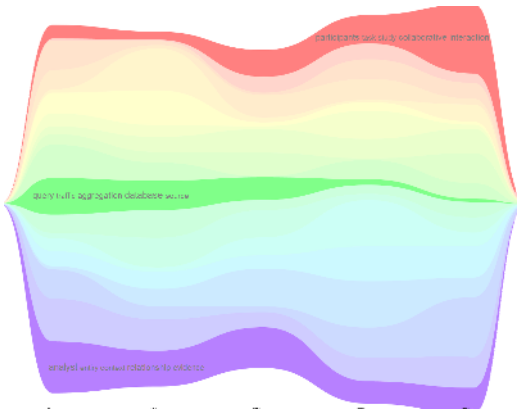

Figure 7: Temporal view with three topics highlighted. Each topic is labeled with its first five keywords. Topic in red: participants task study collaborative interaction; topic in green: query traffic aggregation database source; topic in purple: analyst entity context relationship evidence.

The order of the topics (from top to bottom) is the same as in both the Document Distribution view and the Topic Cloud. We assign the topic colors by interpolating a color spectrum using the normalized distance (Equation 1) among all adjacent topics. As a result, a more similar pair of topics is assigned with colors that are more alike.

Overall, the Temporal view provides a visual summary of how topics of the document collection evolve over time. Beyond the

representation, various interactions are supported within the Temporal view. Selection of a time frame (one vertical time unit) would result in the filtering of all documents published within the selected time frame. Similarly, clicking on an intersection of a topic ribbon and a time frame in the Temporal view will lead to the selection of documents published during the time period with more than 30% probability on the selected topic (Figure 8). Therefore one may identify what documents contribute to the rise of a topic in a certain time period. The view adds richness to ParallelTopics by revealing temporal information hidden in a document collection and allowing users to perform filtering based on time and topic.

3.2.5 Details on Demand

In ParallelTopics, upon selection of any documents, we provide details of the actual text content of the documents of interest. Since any topic models are far from perfect, the function of the detail view is two-fold: first, it provides context for users to develop a deep understanding of a topic and its associated keywords. Second, the detail view helps users to validate the patterns shown in the visualization.

3.2.6 View Coordination and Interactions

Since making sense of a large text corpus may involve the utilization of all four views, coordination among all views is carefully crafted within the ParallelTopics. On the *topic* level, hovering over a topic in any view that involves topic representation would highlight the same topic in other views. For example, if a user hovers the mouse over an axis in the Document Distribution view, the same topic would be highlighted in both Topic Cloud view and Temporal view (Figure 1). Thus the user could quickly synthesize information regarding keywords, document distribution, and temporal trend of the particular topic. In addition, the views are also coordinated by colors, with each topic being the same color in all views.

On the *document* level, selecting any set of documents in views that involve individual document would highlight the same set of documents in other views. For example, brushing in the Document Scatterplot would be immediately reflected in the Document Distribution view, and vice versa. When a user selects a few documents with two prominent topics (mid-range) in the scatterplot, seeing the distributions of these documents helps the user understand their topical combinations.

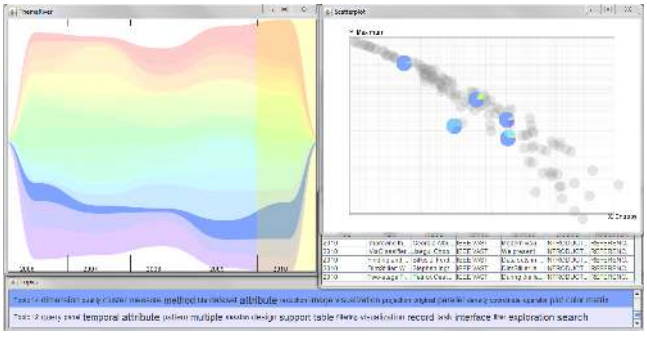

Figure 8: Selection of topic "dimension reduction" and year 2010 in the Temporal view. Documents that focus on this topic and were published in 2010 are shown in other views.

With regard to the *temporal* aspect, filtering documents that were written/published within a certain time period is supported. For instance, clicking on a time frame (one vertical time unit) in the Temporal view would result in the filtering of all documents published within the selected time span. Similarly, clicking on an intersection of a topic ribbon and a time frame in the Temporal view will lead to the selection of documents published during the time period with the topic accounting for a major contribution (more than 30% probability) to those documents (Figure 8). Such selection will be shown in both the Document Distribution view and the Document Scatterplot view. The function allows users to filter documents based on time and topic of interest and then examine the documents published within the selected time frame.

The ParallelTopics supports users to explore and query large document corpora from multiple aspects. Starting with the Topic Cloud, a user could view a summary of the corpus and may identify topics or even keywords of interest. From the Document Distribution view, the user may locate the topic of interest and select documents that focus on this topic by brushing on the topic. The user could then visually identify what other topics the selected set of documents are related to through viewing the distributions in the Document Distribution View and Document Scatterplot. Furthermore, the user could always examine details of the documents upon selection(section 3.2.5). If the user wants to identify interdisciplinary/multidisciplinary publications from the corpus, she is equipped to do so in the Document Scatterplot view by selecting documents in the mid to lower right corner. What's more, if the user is interested in querying the corpus by temporal factor, she may perform selections in the Temporal view though either clicking on one time frame or on an intersection of a certain time frame and a topic. In summary, ParallelTopics employed multiple coordinated views to support interactive exploration of text corpora. Each of the views is designated to address one out of four important questions. As we will see in the next section, different combinations of these questions constitute a versatile set of tasks within specific domains.

4 Case Studies

To evaluate the efficacy of our system in answering the four intended questions, we applied ParallelTopics to exploring and analyzing two text corpora: the scientific proposals awarded by the National Science Foundation and the publications in the IEEE VAST proceedings from 2006 to 2010. We then conducted user evaluations with experts from both domains. For our study with the NSF proposals, we invited a former program manager to use the ParallelTopics system to explore the proposals to evaluate whether the system could assist her in decision-making and award portfolios management. For our study with the VAST publications, 4 researchers in the field of visual analytics used the ParallelTopics to freely explore the publications from the most important venue in the field. In both studies, we performed expert evaluations in which we first provide training to the participants and then asked them to analyze the corresponding dataset using our system. We recorded the feedback from these participants, and conducted a post-study interview to collect their general feedback about the visualizations, the tasks, and the efficacy of the overall system.

4.1 Case Study 1 Analyzing science proposals

In this case study, we first describe the data we collected. We then characterize the targeted domain and present a set of tasks that are summarized based on our conversations with program managers at NSF. Last, we present how ParallelTopic could assist the expert user in solving these tasks.

4.1.1 Data Collection and Preparation

To examine whether ParallelTopics could assist program managers in making funding decisions and managing award portfolios, we first collected the awarded proposals from 2000 to 2010 under the IIS (Information & Intelligent Systems) division, which is part of the CISE (Computer & Information Science & Engineering) directorate. The collection consists of nearly 4,000 awards, with structured data on the *Award Number, Directorate, Division, Program, Program Manager, Principal Investigator,* and *Award Date*; as well as *Abstract* of the proposals, which is in the form of unstructured

text. We processed all collected abstracts with each abstract constituting a single document in the corpus. We removed a list of standard stopwords. This gave us a vocabulary of 334,447 words. We then extracted 30 topics from the corpus using the LDA model.

4.1.2 Domain Characterization

A core part of NSFś mission is to keep the United States at the leading edge of discovery, both by funding research in traditional academic areas, including identifying broader impacts, as well as funding transformative and interdisciplinary research. In order to do the former, the program managers at NSF need to identify appropriate reviewers and panelist to ensure the best possible peer review. In order to effectively perform the latter, the program managers need to identify emerging areas and research topics for funding interdisciplinary and transformative research. In addition to making funding decisions, program managers also need to manage their award portfolios. While the program managers have done a great job in the past, they are in need of new methods to help them due to the rapidly changing nature of science, and the significant increase in the number of proposal submitted [23]. Mapping the high-level mission to actionable items, we designed 3 tasks that are related to decision-making and award portfolio analysis. Task 1 focuses on dividing new proposal submissions into groups based on their topics. This task requires understanding the *major topics of the text corpus (Q1)*, and filtering sub document collections based on their *characteristics over topics (Q2)*. Task 2 is to identify appropriate reviewers for the proposal submissions, which further involves knowing *whether a submission is related to multiple topics (Q3)* in order to gather the right expertise. Last, Task 3 focuses on the temporal aspect of the award portfolios which involves *discovering the topical trend over time (Q4)*.

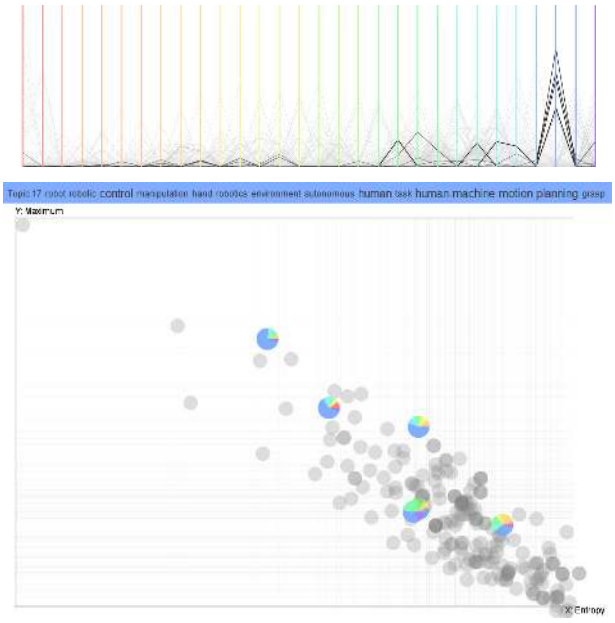

Figure 9: Exploration scenario. Top: Selection of documents on the topic “robotics”. Bottom: Pie glyphs in the Document Scatterplot show the number of topics for each selected document respectively.

4.1.3 Expert Evaluation

Since program managers at NSF are extremely busy, we invited a former NSF program manager for our expert evaluation. The participant has two years of experience working as a program manager at NSF. At the beginning of the evaluation, we spent 30 minutes demonstrating the system design and functionality of each visualization. Then we asked the participant to perform the following three tasks using the ParallelTopics system.

Task1: To group 200 newly submitted proposals based on topics [1]

Starting with the Topic Cloud, the participant quickly scanned the extracted topics to gain an overview of the newly submitted proposals. Since the participant was responsible for proposals in the areas of robotics and computer vision, she quickly focused her attention on these two topics. Upon selection of the proposals that focus on the topic regarding robotics (figure 9), the participant quickly glanced over the titles in the detail view to validate their relevancy. While the participant was making sure that each selected proposal is relevant, she also noticed that the positions of the proposals are scattered in the Document Scatterplot. Since the proposals in the lower right positions are more likely to contain two or more topics, the participant was interested in knowing what other topics these proposals relate to. Through further filtering the proposals that appear to be more interdisciplinary in the Document Scatterplot, the participant found that they are related to other fields such as neuroscience and social communication.

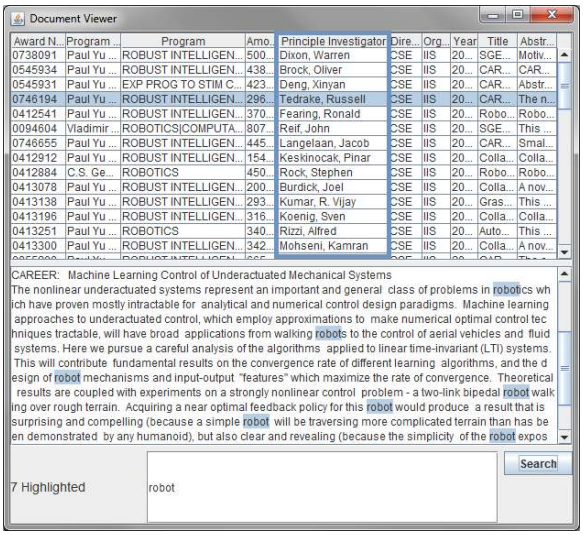

Figure 10: Upon selecting the relevant document in the Document Distribution view, the Detail view is invoked so that the program manager can look for previously awarded PIs.

Task2: To identify appropriate reviewers For the purpose of identifying reviewers, the participant first wanted to roughly divide the proposals into groups. Based on the initial exploration, the participant concluded that there are roughly two groups of proposals: one group that focus on the core of robotics area, and the other that utilized body of knowledge from other fields such as neuroscience and social communication. To identify reviewers for the two groups of proposals, the participant would like to find PIs from previously awarded proposals. Through examining the historic data, the program manager located the topic regarding robotics in the Document Distribution view. She then brushed the top range of the axis to select proposals pertinent to the topic. Finally, the participant turned to the detail view (figure 10) to look for PIs that were previously awarded in the robotics area. For interdisciplinary proposals in group2, the participant went through similar processes to identify

[1] Since awards that were not accepted are considered proprietary, we collected 200 awarded proposals in the year 2010 to mimic the newly submitted proposals. The proposals awarded from 2000 to 2009 serve as historic data.

additional experts from other related field (e.g. neuroscience) to serve on the review panel to ensure the best possible peer review.

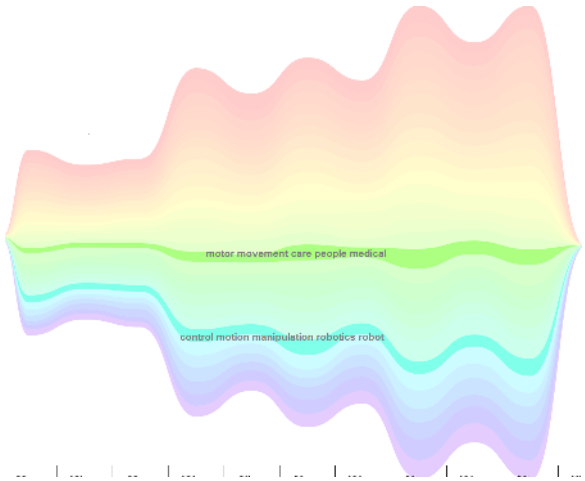

Figure 11: Although the total number of proposal grew continuously over the years, the awarded proposals regarding topic "robotics" remained steady (light blue). In contrast, more proposals related to "using interfaces to help people with impairment" were awarded over the years (green).

Task3: Analyzing temporal trend of award portfolio On a portfolio level, the former program manager was interested in seeing the temporal trend of the areas she is in charge with over years. Through exploring the Temporal View, the participant discovered that the trend of awarded proposals in the field of robotics is steady, although the overall number of proposal awarded grew during year 2006 and 2009. Unlike the steady trend of robotics, the number of awarded proposals on the topic of "using technology to help people with disability" grew over the years (Figure 11). The former program manager commented that this view is valuable to her since it enabled her to see funding trends regarding different topics that are otherwise hard to discover.

In summary, the participant thought each view in the ParallelTopics system is well designed with a clear purpose. She commented that the tool could play a facilitating role in a program manager's workflow. In particular, she liked the fact that our tool could automatically suggest proposals that are more interdisciplinary since this was difficult to judge traditionally. She also liked coordination between views, which helped her to quickly synthesize information from different aspects of the same corpus.

During the post interview, the participant suggested a few potential areas of improvements. Specifically, the participant would like the ParallelTopics system to provide information on identifying conflict of interest when assigning reviewers. On a portfolio management level, she would also like to see funding trend over other aspect such as geographic regions.

4.2 Case Study 2 Analyzing VAST conference proceedings

As the field of visual analytics mature, it is helpful to review how the field has evolved. One means to approach this problem is to analyze the publications that have been accepted by the most important venue in Visual Analytics. In this case study, we recruited 4 researchers to explore articles published in the VAST conference/symposium since the field began in 2006. Since all users were familiar with the field of visual analytics, we wanted to encourage free exploration as opposed to following well-structured tasks. After the evaluation, we categorized the findings from the participants into two groups: discovering causal relationship between temporal evolution of topics (*Q1,Q4*) and funding sources; learning about interesting subfields in the realm of visual analytics (*Q1, Q2, Q3*).

4.2.1 Data Collection and Preparation

We first collected all articles published in the VAST conference/symposium from 2006 to 2010. A total number of 123 publications were collected. We then parsed each publication into fields including *Title, Author, Year Published, Abstract, Body*, and *References*. We performed topic modeling on the full body of each paper (from introduction to conclusion) with each paper constituting a document in the corpus. Removing standard stopwords left us a vocabulary of 317,315 words. Based on our tally of different tracks for every VAST conference, we extracted 19 topics from the corpus.

4.2.2 User Evaluation

Among the 4 researchers we recruited, 2 are senior researchers in the field of visual analytics and the other two are Ph.D. students with visual analytics as their main research interest. In this evaluation, we provided all participants a high-level task and encouraged more free exploration. After introducing the ParallelTopics system, we asked each participant to identify core topics within the field and how the field has evolved over the course of last 5 years. We roughly categorized the usage patterns into two groups: identifying rising/falling topics and using the system as an educational tool.

Identifying rising/falling topics After glancing through all topics in the Topic Cloud, one senior researcher commented that the topics conform well to the paper tracks from the VAST conferences. When viewing the temporal trend of each topic, the participant noticed a few clear rising and falling patterns. For instance, the topic on video news analysis attracted lots of interest at the beginning, but the interest quickly diminished over the years (Figure 12, yellow ribbon). He also noticed a similar trend on the topic regarding network traffic monitoring and analysis (Figure 12, green ribbon). Associating the patterns with his knowledge, the participant explained the trends as when the field began, the areas of interest were guided by DHS which is the primary funding source at the time. Next, the participant turned to the rising patterns which indicate interests in those topics grew over the years. In particular, both topic trend and uncertainty analysis and topic dimensionality analysis and reduction attracted more interests since year 2008 (Figure 13). Again associating the patterns with his own knowledge, the participant commented that this is likely the outcome of the FODAVA (Foundations of Data and Visual Analytics) program introduced by NSF and DHS jointly.

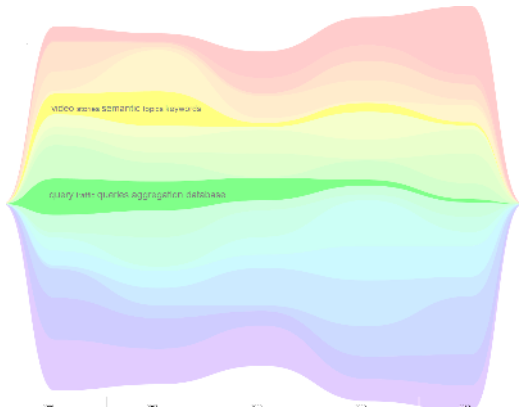

Figure 12: Falling topics identified by the expert user.

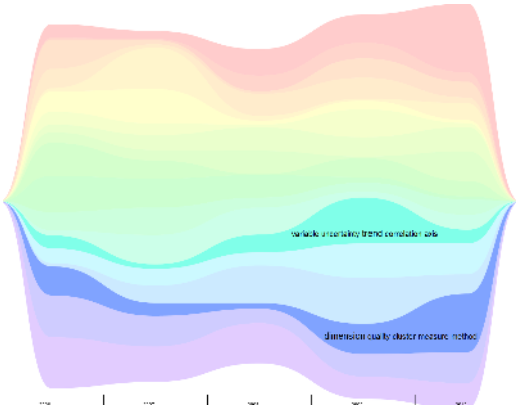

Figure 13: Rising topics identified by the expert user.

Learning about the field of visual analytics The other senior researcher who was teaching a visual analytics course at the time commented that he can see the tool being useful for his class. Students could explore all VAST publications and identify papers that related to topics of interest for course presentation. Similarly, another participant wanted to see what has been done on text analysis in the field of visual analytics. He first located the topic and then selected publications that ranked high on this topic in the Document Distribution view. He quickly glanced through the paper titles in the detailed view and validated all papers selected were of his interest. He also noticed that some papers within the selection appear related to other topics such as entity extraction and database queries (figure 14). After the study, he asked for a screen capture of the detailed view so that he could look for the papers he identified during the study.

In summary, the participants considered the ParallelTopics system useful in helping them explore the evolution of the field of visual analytics and identify publications for further investigation based on their own interest.

5 Discussion and Future Work

ParallelTopics utilizes LDA as the automatic text summarization method. Since the debut of LDA in 2003, a lot of research efforts have been devoted to improving the initial generative topic model. As a result, variations of LDA extensions such as the correlated topic model (CTM) [8], the dynamic topic model [7] and OnlineLDA [17] have been developed to address limitations of LDA. Given the large number of topic models available, the best model for summarizing text corpora may depend on the characteristics of the corpora. However, the visual metaphors built in the ParallelTopics that are rather invariant to the underlying probabilistic models. In other words, the visual metaphors could easily adapt to many variations of probabilistic topic models.

One issue worth noting when using topic models to summarize text corpora is the decision on the number of topics that best describe the corpora. Since there has been no explicit way to quantitatively evaluate the topics , researchers have employed a variety of metrics of model fit, such as perplexity of a probility model or held-out likelihood [10]. However, such measures do not indicate how interpretable the latent space is [10]. Therefore, human involvement is still the best way to ensure the generation of semantically meaningful topics. ParallelTopic allows users to quickly glance over the extracted topics and more importantly, to visualize how the documents distribute across topics. If one of the topics only attracts very few documents with low probability on the topic, it is likely to be an outcome of too many topics that overfit the corpus. Therefore the user may decide to extract fewer topics based on such

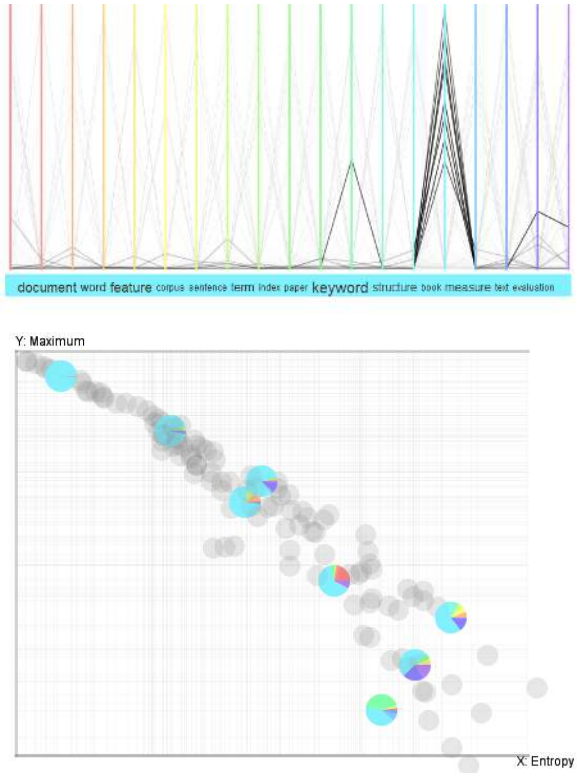

Figure 14: Top: Selection of documents on the topic “document analysis”. Bottom: Pie glyphs in the Document Scatterplot show the number of topics for selected documents. The documents on the bottom right contain mutiple topics.

observation.

In terms of the scalability of ParallelTopics, large number of topics and documents will result in visual clutter. In the future, we plan to reduce such clutter in two ways: from the computational side, we plan to build a hierarchical structure for the topics and allow users to drill down to lower-level topic hierarchy on demand; on the visual side, we wish to group documents based on their similarity and select an exemplar for each group to reduce document clutter within the visualization.

6 Conclusion

In this paper, we present a novel visual analytics system, ParallelTopics, to enable users to interactively explore large text corpora. The ParallelTopics system utilizes a probabilistic topic model to summarize the text corpora and highlights the probabilistic nature of each document across topics. Such representation allows users to interactively identify multi-topic documents, which os difficult to do with existing visual analytics systems. In addition, the ParallelTopics system presents the probabilistic distributions of documents from a temporal perspective, which supports the exploration of topical trends. Synthesizing information from multiple views provided by ParallelTopics, a user could gain a deep understanding of the otherwise unstructured text collection. Our evaluation with expert users on two different corpora demonstrates that ParallelTopic effectively assists users in multiple text analysis tasks through addressing four fundamental questions regarding a text corpus.

Acknowledgements

This work is supported by the National Science Foundation under award number 0915528.

REFERENCES

[1] Pacific northwest national laboratory. http://in-spire.pnl.gov/.

[2] Committee on facilitating interdisciplinary research committee on science policy. Facilitating interdisciplinary research, National Academies. Washington: National Academt Press 2004.

[3] L. Alsumait, P. Wang, C. Domeniconi, and D. Barbar. *Embedding Semantics in LDA Topic Models*. John Wiley & Sons, Ltd, Chichester, UK, 2010.

[4] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann. The infosky visual explorer: exploiting hierarchical structure and document similarities. *Information Visualization*, 1:166–181, December 2002.

[5] D. Blei, L. Carin, and D. Dunson. Probabilistic topic models. *Signal Processing Magazine, IEEE*, 27(6):55 –65, 2010.

[6] D. Blei and J. Lafferty. *Text Mining: Theory and Applications*, chapter Topic Models. Taylor and Francis, 2009.

[7] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[8] D. M. Blei and J. D. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35, June 2007.

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

[10] 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.

[11] Y. Chen, L. Wang, M. Dong, and J. Hua. Exemplar-based visualization of large document corpus (infovis2009-1115). *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1161 –1168, 2009.

[12] P. Crossno, D. Dunlavy, and T. Shead. Lsaview: A tool for visual exploration of latent semantic modeling. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 83 – 90, 2009.

[13] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1017 –1026, 2010.

[14] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235, Apr. 2004.

[15] T. L. Griffiths, J. B. Tenenbaum, and M. Steyvers. Topics in semantic representation. *Psychological Review*, 114:2007, 2007.

[16] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, INFOVIS '00, pages 115–, Washington, DC, USA, 2000. IEEE Computer Society.

[17] M. D. Hoffman, D. M. Blei, and F. Bach. Online Learning for Latent Dirichlet Allocation. In *Neural Information Processing Systems*, 2010.

[18] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1:69–91, 1985. 10.1007/BF01898350.

[19] T. Iwata, T. Yamada, and N. Ueda. Probabilistic latent semantic visualization: topic model for visualizing documents. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 363–371, New York, NY, USA, 2008. ACM.

[20] J. B. Kruskal. Multidimensional scaling. *Sage University Paper series on Quantitative Application in the Social Sciences*, pages 07–011, 1978.

[21] T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211 – 240, 1997.

[22] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. Interactive, topic-based visual text summarization and analysis. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 543–552, New York, NY, USA, 2009. ACM.

[23] NSF. Discovery in a research portfolio: Tools for structuring, analyzing, visualizing and interacting with proposal and award portfolios.

[24] P. Oesterling, G. Scheuermann, S. Teresniak, G. Heyer, S. Koch, T. Ertl, and G. Weber. Two-stage framework for a topology-based projection and visualization of classified document collections. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 91 –98, 2010.

[25] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[26] D. Ramage and E. Rosen. Stanford Topic Modeling Toolbox, http://nlp.stanford.edu/software/tmt/tmt-0.3/.

[27] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24:513–523, August 1988.

[28] G. Salton, A. Wong, and C. S. Yang. *A vector space model for automatic indexing*, pages 273–280. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[29] 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.

[30] 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.

[31] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM.

[32] W. Zhu and C. Chen. Storylines: Visual exploration and analysis in latent semantic spaces. *Computers & Graphics*, 31(3):338 – 349, 2007.