Exploring Topical Lead-Lag across Corpora

Shixia Liu, Yang Chen, Hao Wei, Jing Yang, Kun Zhou, and Steven M. Drucker

Abstract—Identifying which text corpus leads in the context of a topic presents a great challenge of considerable interest to researchers. Recent research into lead-lag analysis has mainly focused on estimating the overall leads and lags between two corpora. However, real-world applications have a dire need to understand lead-lag patterns both globally and locally. In this paper, we introduce *TextPioneer*, an interactive visual analytics tool for investigating lead-lag across corpora from the global level to the local level. In particular, we extend an existing lead-lag analysis approach to derive two-level results. To convey multiple perspectives of the results, we have designed two visualizations, a novel hybrid tree visualization that couples a radial space-filling tree with a node-link diagram and a twisted-ladder-like visualization. We have applied our method to several corpora and the evaluation shows promise, especially in support of text comparison at different levels of detail.

Index Terms—Text visualization, Lead-lag analysis, Tree visualization, Adaptive focus + context.

1 Introduction

In many applications, it is desirable to identify which text corpus (lead) is followed by others (lags) regarding a specific topic. For example, given a set of publications and grant proposals, as well as a topic of interest (e.g., hardware-based rendering), a funding officer wants to know whether a research trend in publications leads that of the grant proposals on this specific topic or vice versa. S/he relies on the lead-lag relationships between the publications and proposals to adjust funding allocations or to initiate new programs. Another example is information diffusion among social media. Businesses and organizations are interested in knowing which social media outlet first publishes messages pertaining to a specific event (e.g., the financial crisis) and how the event is propagated across different outlets. This has an important impact on their product, brand, and policy communication strategies, as well as on the reaction time for critical events such as natural disasters. Accordingly, there is an increasing need to create algorithms and visualizations to help understand the lead-lag relationships across corpora at both a global level (across all time), and at the local level (for individual time points).

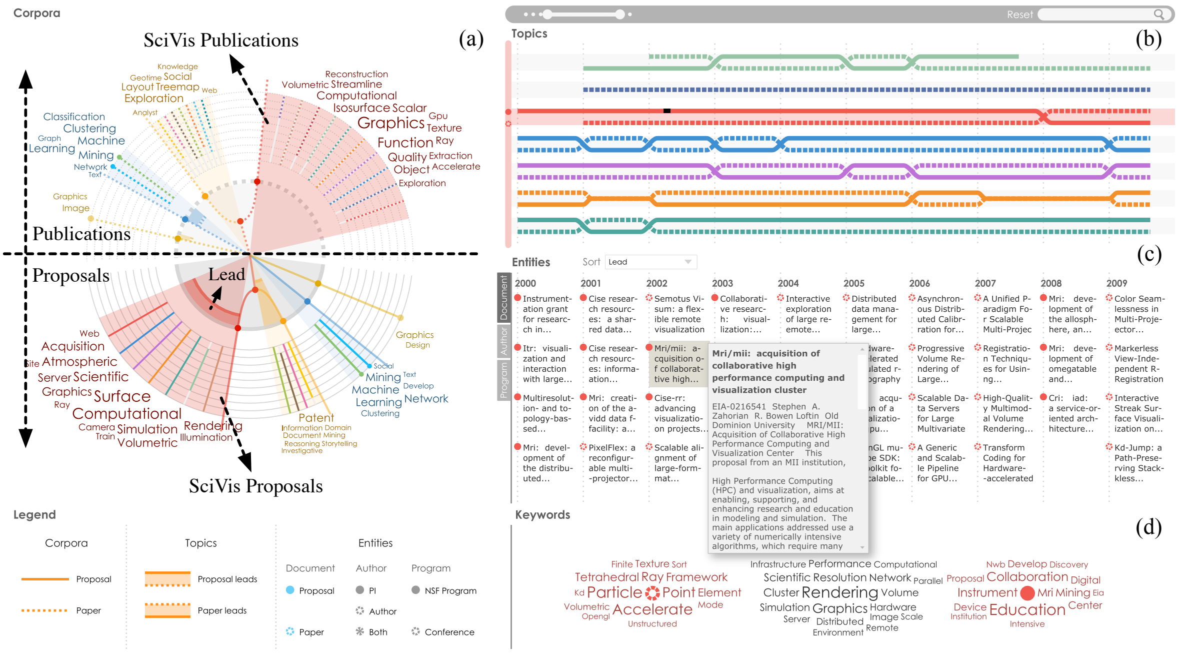
Existing solutions for lead-lag analysis come mainly from the field of text mining. Researchers in the field have made attempts to derive overall lead-lag relationships between corpora using latent Dirichlet allocation (LDA) and time series analysis [1], [2]. However, they do not always fulfill the complex tasks in real-world applications, where users need to not only examine global patterns, but also analyze local lead-lag changes regarding a specific topic over time.

As an example to motivate our work, we show how lead-lag relationships can help a funding officer to determine whether some specific existing areas in the field of Scientific Visualization need to be funded or whether to initiate some specific new programs. To do this, she examines two corpus, one containing papers and another consisting of proposals.

As shown in Fig. 1(a), the corpus on the top is the publication corpus (dotted lines), and the one on the bottom is the proposal corpus (solid lines). SciVis is exemplified by keywords such as “Surface,” “Volumetric,” and “Isosurface.” This area is led by proposals slightly (encoded by a filled portion). She then examine the local lead-lag relationship changes of each topic pair in SciVis. In Fig. 1(b), there are 7 total topics and 6 topic pairs in the *topics view* (the single blue topic only exists in the paper corpus and not in the proposal corpus). For each topic pair, the one on the top is the lead. Most of the topic pairs change their lead-lag relationships over time. One pattern that seems to differ from the other patterns is the red topic which shows that proposals primarily led papers from the years 2000 to 2008. After examining its content in the *keywords view*, the officer find significant keywords such as “Rendering,” “Hardware,” and “Resolution” (Fig. 1(d)). This pattern may be due to the fact that before conducting research on hardware, researchers need grants for the acquisition of hardware equipment. Examining the proposals from the *entities view* confirmed this hypothesis; many of them are grants relating to hardware equipment acquisition. For instance, one proposal in 2002 (Fig. 1(c)) is about the acquisition of high performance computing and visualization clusters. The single blue topic also attracts our attention, which is a topic from publications without a pair in the proposals. By examining the content, she discover that it is a well-established research topic, “geometric modeling and rendering,” so the NSF officer can focus funding allocations on other needed areas.

Motivated by the above example, we are interested in analyzing the lead-lag patterns at both the global and local levels in this work. There are two major technical challenges. One is to model lead-lag across corpora globally and locally.

- *S. Liu and S. Drucker are with Microsoft Research.*
  E-mail: shliu@microsoft.com,sdrucker@microsoft.com
- *Y. Chen and J. Yang are with the Department of Computer Science, University of North Carolina at Charlotte, USA. E-mail: ychen61@uncc.edu, jing.yang@uncc.edu*
- *H. Wei and K. Zhou are with the Department of Computer Science, Zhejiang University, China. E-mail: v-hawe@microsoft.com, kunzhou@acm.org*



Mri/mii: acquisition of collaborative high performance computing and visualization cluster

EIA-0216541 Stephen A. Zahorian R. Bowen Loftin Old Dominion University MRI/MII: Acquisition of Collaborative High Performance Computing and Visualization Center This proposal from an MII institution,

High Performance Computing (HPC) and visualization, aims at enabling, supporting, and enhancing research and education in modeling and simulation. The main applications addressed use a variety of numerically intensive algorithms, which require many

Fig. 1. Lead-lag analysis results between the NSF proposals and academic publications in the following research areas: visualization, data mining, and computer graphics.

The lead-lag relationships in different corpora often change over time. Therefore, it is difficult to model them using existing global-level approaches that only aim to extract the overall leads. Another challenge is to design an intuitive and consistent visualization that allows users to better understand lead-lag patterns, as well as to investigate the major causes that lead to these patterns. The lead-lag analysis module provides two-level results, from the global level to the local level. Such results usually need several visualizations to convey them. It is therefore preferred to design a coherent visualization mechanism that can be tightly integrated with the analysis module.

To address these challenges, we have developed *TextPioneer*. A demo video is available at http://research.microsoft.com/en-us/um/people/shliu/TextPoineer-final.wmv. The major contributions of this work are:

- **An interactive visual analytics tool** that tightly integrates interactive visualization with lead-lag analysis to help users to better understand lead-lag relationships across corpora both globally and locally. It enables coordination between the views such as brushing and linking, so that users can explore the analysis results at different levels.
- **A two-phase analysis mechanism** that seamlessly accomplishes lead-lag analysis at both the global and local levels. The global analysis starts by discovering evolving topics across corpora using an evolutionary hierarchical Dirichlet processes (HDP) model [3]. Then the overall topical lead values are computed. The local analysis aims to automatically identify the lead-lag changes among the topics over time.
- **A coherent visualization mechanism** that encodes the two-level analysis results. To illustrate the global lead-lag among corpora, we have developed a novel hybrid tree visualization that couples a radial space-filling tree visualization with a node-link diagram (Fig. 7(a)). As a user selects the topics of interest from the hybrid tree, a twisted-ladder-like visualization (Fig. 7(b) and Fig. 7(c)) is then provided to convey their local changes and document alignment over time. Here, document alignment indicates that highly similar document pair across corpora are connected to help users track why the lead-lag relationships change over time.

2 Related Work

2.1 Lead-Lag Analysis

Much effort has been made to discover topical leads and lags across corpora. These approaches can be classified into two categories based on the types of the topics: named-entity-based approaches and topic-based approaches.

Named-entity-based approaches treat named entities or distinctive phrases as topics and then study their overall lead-lag behaviors. For example, Lloyd et al. [4] studied the lead-lag relationships between blogs and the news media by analyzing the reference frequency time series of named entities that appear often in both sources. *MemeTracker* [5] focused on providing a coherent representation of the news cycle, as well as tracking the evolution patterns of short phrases (or memes) across the news media and blogs.

Although named entities or distinctive phrases act as signatures of topics, they do not provide an adequate representation of semantic topics since the contextual information is lost. To tackle this problem, lead-lag analysis based on the topic model has been developed. Menezes et al. [6] proposed a likelihood-based probabilistic approach to estimate the temporal relationships between blogs. In this approach, they defined the precursor score of blog **A** with respect to blog

B as the probability that **A** introduces a new topic before **B**. Gerrish and Blei [7] proposed a dynamic topic model to identify the most influential documents conditioned on the topics. This model uses the content change of documents to measure the importance of individual documents within a collection. Zhang et al. [3] propose evolutionary hierarchical Dirichlet processes to handle multiple text sources. This method aimed to analyze and compare topic evolution patterns across different corpora. The evolutionary HDP is formulated as a set of hierarchical Dirichlet processes by adding time dependencies to the adjacent time points.

Compared with the above methods, our approach not only tracks topic evolution within and across corpora, but also extracts the global and local lead-lag analysis results. Furthermore, we leverage visualization techniques to allow users to interactively explore the flow of ideas at the corpus, topic, and document levels. The multi-scale, interactive approach helps users deeply understand the lead-lag patterns and what contributes to the patterns.

Similar to our analysis method, Shi et al. [2] computed the overall topical lead-lag based on purely textual and time-stamp information using LDA [8] and time series analysis. Recently, Nallapati et al. [1] used a simple TF-IDF based nearest-neighbors approach to estimate the topical leads between two corpora. They also proposed LeadLag LDA to identify the idea flow on specific topics. Shi and Nallapati et al.'s work focused only on inferring global leads and lags. In addition to the global patterns, our approach also extracts local lead-lag relationships over time. Furthermore, we also leverage interactive visualization to illustrate the lead-lag pattern from the global level to the local level and explore the major factors that lead to them.

2.2 Visual Comparison of Hierarchical Data

Many visualization methods have been developed to compare hierarchically organized data. For example, *TreeJuxtaposer* [9] compared large trees side-by-side and highlighted similar nodes and sub-tree structures using colors. To support multiple object comparison, Bremm et al. [10] designed a system that provides a set of interlinked hierarchy views for comparing multiple trees. Each view represents a sub-tree and shows both global and local similarities to a selected reference tree. A schema mapper visualization was proposed [11] to explain one-to-one mappings between two XML schemas. Holten and van Wijk [12] visualized the relationships between matching sub-hierarchies using hierarchical edge bundles. With these bundles, users can easily identify elements that are unique to each hierarchy.

On the other hand, Graham and Kennedy [13] merged multiple trees into a unified structure and visualized it using a directed acyclic graph, where overlaps and differences between groups of trees and individual tree are revealed. Tu and Shen [14] constructed a union treemap for two trees that are being compared. Corresponding items in the trees are mapped to a single item and their attribute differences are shown by blending their attribute colors. The structural differences are visualized using special colors. Isenberg et al. [15] presented a tree comparison system for co-located collaboration where semi-transparent trees from individuals overlapped, with the best matched nodes overlaid. Similar to the above visual tree comparison methods, *TextPioneer* also visually compares the topic hierarchies of different corpora. However, existing approaches do not apply to this problem since additional lead-lag relationships need to be encoded together with the tree structure. As a result, we have designed a hybrid tree visualization to enable the comparison of several corpora with hierarchies simultaneously. Furthermore, *TextPioneer* uses a twisted-ladder-like visualization to illustrate local lead-lag relationships over time.

2.3 Visual Topic Evolution

Much research has been performed to analyze temporally evolving topics in text corpora. ThemeRiver [16] visually depicted how the keyword strengths change over time in a text corpus using a river metaphor. TIARA [17], [18], [19] employed an LDA-based topic model to analyze a large text corpus and depict how the topics evolve over time using an enhanced stacked graph. TextFlow [20], [21] was developed to visually convey topic merging/splitting relationships over time. EventRiver [22] assumed that clusters of news articles with relevant content are adjacent in time and can be mapped to events. Thus this method automatically detected such clusters and visually depicted their strength over time to reveal the impact of these events. Outflow [23] merged multiple event sequences into a graph. With this abstraction, it enabled more scalable timeline visualizations. More recently, storyline visualization [24], [25] is developed to depict dynamic social interactions in a story or an event, over time.

The above approaches focus on topic exploration on individual text corpora. To the best of our knowledge, our work is among the first to support visual exploration of the lead-lag relationships between/among corpora.

3 System Overview

TextPioneer consists of three components: a document processor, a lead-lag analyzer, and a lead-lag visualizer (Fig. 2). Given several text corpora, the document processor first extracts the document body text and relevant metadata such as author and time stamp information for each document. The output is then fed to the lead-lag analyzer, which identifies a set of hierarchical topics, global and local lead-lag relationships, and a group of influential documents and authors. Next, the visualizer transforms the two-level results into a consistent visualization, which contains two major visual components. Specifically, a hybrid tree visualization is designed to encode the overall lead-lag at the corpus and topic levels, and a twisted-ladder-like visualization is developed to illustrate the local lead-lag relationships and their document alignment over time.

4 Topic-based Lead-lag Analysis

To help users analyze lead-lag relationships across corpora, we have developed an approach based on the evolutionary

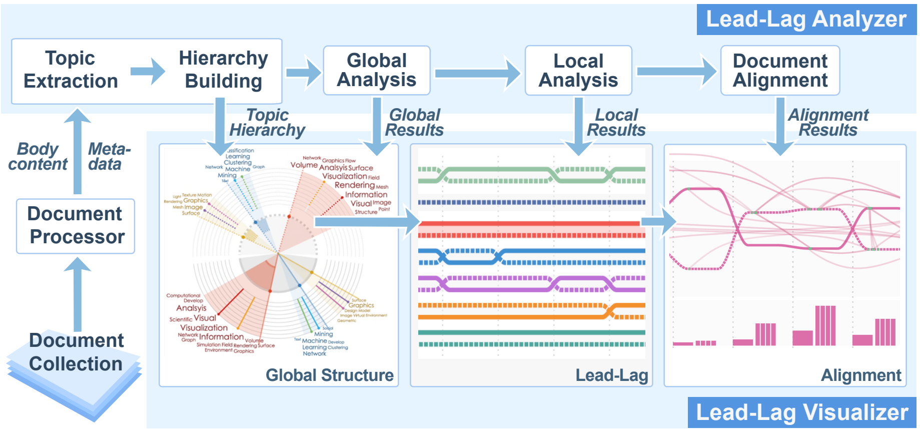


Fig. 2. *TextPioneer* overview.

hierarchical Dirichlet process (HDP) model [3]. In this approach, we first leverage evolutionary HDP to extract topics from multiple corpora and then derive a topic hierarchy to organize them. Next, we estimate the leads and lags globally and locally.

4.1 Topic Extraction and Hierarchy Building

Research in the text mining community has shown that the HDP model is very effective in performing topic extraction for two major reasons: 1) it effectively models topics across multiple corpora; 2) it automatically determines the optimal topic number [3], [26], [27]. Therefore, we adopt the evolutionary HDP clustering model [3] to learn evolving topics. With this clustering model, we extract a set of topics shared by multiple corpora and assign each document to the topic with the highest probability value. Typically, a topic represents the thematic content common to a set of text documents across corpora. It is characterized by a distribution over a set of keywords. Each keyword has a probability value to measure its likelihood of appearing in the related topic.

In real world applications, it is very common to generate dozens or even hundreds of topics. If we display all of them, it would cause visual clutter. To solve this problem, we build a topic hierarchy to organize the extracted topics. With it, users can easily examine the leads and lags at different granularities, including topics and topic clusters. In TextPioneer, we adopt the state-of-the-art method, rose-tree-based hierarchical clustering [28]. It builds a multi-branch topic hierarchy based on the topic keywords.

4.2 Lead-Lag Analysis

TextPioneer aims to analyze the overall lead-lag relationships across corpora and their changes over time. We therefore model such temporal relationships across corpora globally (across all time) and locally (at each time). Here we take two corpora as an example to illustrate the basic idea. We first introduce some preliminary definitions:

Local lead: Given two corpora A and B, if the content of A at a given time point tends to be more similar to the future of B than its past in the context of a given topic, then A is the local lead at this specific time point.

Global lead: If the average lead value (at all time points) of A is greater than B, then A is the global lead.

4.2.1 *Global Analysis*

Our approach to modeling global lead is inspired by **Lead-Lag LDA** [1]. It assumes that novel ideas are developed and diffused through the creation and propagation of new or newly emphasized keywords. As a result, the lead-lag relationships between corpora (in the context of a given topic) can be estimated by word usage across them. Particularly, our approach consists of two steps: nearest neighbor finding and lead-lag inference.

In nearest neighbor finding, a **TF-IDF** based nearest-neighbor method is adopted. Given two text corpora A and B, let A be the reference corpus, and B be the comparative corpus. For each document d_A in topic T of A, the algorithm first retrieves the most similar documents in B using the **TF-IDF** measure. Here, we use a predetermined document similarity threshold τ to select the significant document matches.

In the next step, lead-lag inference, we compute the lead value of a corpus regarding a given topic. For each document d_A in the given topic T of A, the lead value is set as the weighted average of the time differences regarding its nearest neighbors in the same topic of B, which is denoted by $\mathcal{N}(d_A)$. As in **LeadLag LDA** [1], we also adopt the similarity values between d_A and its similar documents in $\mathcal{N}(d_A)$ as the weights for computing the lead value. This is an application-independent criterion. Application-specific information such as paper/proposal metadata, can also be leveraged to determine the lead corpora with respect to a specific topic. For example, if two documents in one topic are written by the same person, then they are much more correlated. Specifically, in the paper/proposal application, the similarity between d_A and $d_B \in \mathcal{N}(d_A)$ is defined as:

$$w_s(d_A, d_B, T) = w_t + w_m \tag{1}$$

where w_t is the **TF-IDF** similarity score, w_m is the metadata weight (a constant) to emphasize the documents with the same metadata. In our implementation, we set it as the author weight.

Furthermore, we also consider the lead impact of each document in $\mathcal{N}(d_A)$ to emphasize the documents that lead many other documents. It is measured by the number of documents that are led by it in T of both corpora.

$$w_l(d_B,T) = \log(N(d|d \in A \bigcup B \text{ \& } t(d) > t(d_B) \\ \text{\& } w_s(d,d_B,T) > \tau)) \qquad (2)$$

where $N(*)$ counts the number of documents that are correlated to d_B but occur after it, $t(d)$ is the time stamp of d, and τ is the document similarity threshold to filter out insignificant document matches.

Accordingly, the expected lead value of document d_A with respect to B in the specific topic is given by

$$Lead(d_A,T) = \frac{\sum\limits_{d_B \in \mathscr{N}(d_A)} w_s(d_A,d_B,T) \cdot w_l(d_B,T) \cdot (t(d_B) - t(d_A))}{\sum\limits_{d_B \in \mathscr{N}(d_A)} w_s(d_A,d_B,T) \cdot w_l(d_B,T)} \qquad (3)$$

The mean lead of corpus A with respect to B is then determined by averaging the lead values of all documents in the given topic of A. In addition to the mean lead that treats each document equally, we can also formulate a weighted lead for the given topic in A. This is achieved by considering the document influence degree of d_A on B, which is estimated by the number of nearest neighbors in B.

$$w_i(d_A,T) = \log(N(\mathscr{N}(d_A))) \qquad (4)$$

Accordingly, the weighted lead of A is defined by

$$Lead(A,B,T) = \frac{\sum_{d_A \in A} w_i(d_A,T) \cdot Lead(d_A,T)}{\sum_{d_A \in A} w_i(d_A,T)} \qquad (5)$$

In the above discussion, we use two corpora as an example for simplicity. Our approach can be easily extended to multiple corpora by comparing each corpus with the reference and then ordering them according to their lead values. Here, the lead-lag value is measured by the nearest neighbors in the comparative corpus, so the lead value may change when the reference is changed. One way to eliminate such bias is to treat each corpus to be compared as a reference and then average the lead values derived from different references.

4.2.2 Local Analysis

Topical lead-lag relationships among corpora often change as a topic evolves over time. To the best of our knowledge, no systematic research has been conducted to address this problem. In this work, we mainly focus on modeling the local lead-lag change at the topic level. We denote the same topic T_i in corpora A and B by $T_i(A)$ and $T_i(B)$, respectively.

One straightforward way to model the local lead-lag on T_i is a bottom-up method based on paired segments. A segment is a set of documents in a continuous time period. This method first uniformly segments the documents in $T_i(A)$ or $T_i(B)$ over time (e.g., every two years). We smooth the lead estimation by partially overlapping adjacent segments. Then we pair the segments of different corpora by their time range. The segments with the same time range are grouped together. Next, for each pair, we compute the lead values of the segments in $T_i(A)$ and $T_i(B)$. Finally, in one corpus, we merge the adjacent segments that have the same lead relationship. This method is very intuitive. However, it is a greedy method, and by no means guarantees the global optimum.

To address this issue, we propose a competition-based method for estimating the lead-lag change over time. First, for $T_i(A)$, we regard the documents that belong to each time point t as a corpus $T_i(A)^t$, and $T_i(B)$ as another corpus. Then we use Eq. (5) to compute the lead value of $T_i(A)^t$ with respect to $T_i(B)$. We also perform the same computation for $T_i(B)$. Next, we compare the lead values of $T_i(A)$ and $T_i(B)$ at the same time point. The corpus with a larger value is believed to have more novel ideas and is considered to be the leader at that time.

4.3 Evaluation

To evaluate the effectiveness of the proposed lead-lag analysis algorithm, we conducted experiments on labeled paper/proposal datasets. The key computation step of both the global and local analysis is to find the matched document pairs from two corpora and compare their lead-lag relationships. Therefore the experiments aimed to compare the learned lead-lag relationships of document pairs with the labeled ground truth. In our experiments, we adopted two datasets: a visualization (VIS) dataset and a data mining (DM) dataset. Each dataset contains a publication corpus and a proposal corpus. The proposal corpus is collected by randomly selecting NSF grant proposals from the research area of visualization or data mining. To collect the corresponding publications, we extracted the abstracts of related papers from the homepage of each selected proposal on the NSF website. We also added some documents as noise data, such as proposals without related papers or papers without related proposals. After collecting the documents, we asked two domain experts to manually label all the related proposal/paper pairs as either lead or lag.

The inter-agreement between two annotators is 91.5%. After selecting the document pairs that they reach agreement on labels, we got two datasets. The first dataset consisted of 42 proposals, 96 papers from the area of visualization, and 184 paper/proposal pairs with lead values. The second contained 69 proposals, 108 papers from the area of data mining, and 248 paper/proposal pairs.

We compared the lead-lag results of our approach with the labeled results in the two datasets and evaluated the comparison results by precision and recall values. In our experiments, we also varied the values of the author weight w_m and similarity threshold τ to determine better values. As shown in Fig. 3 and Fig. 4, our approach achieved a strong performance when $w_m = 0.1$ and $\tau = 0.1$ for the VIS dataset, and $w_m = 0.1$ and $\tau = 0.11$ for the DM dataset. We further evaluated the influence of the author weight on precision and recall. As shown in Fig. 5, the precision, recall, and F-measure values converge when the author weight is great than or equal to 0.1. This demonstrated that $w_m = 0.1$ is reasonable for the application. We also compared our results with those of the TF-IDF-based baseline on the two datasets. As shown in Fig. 6, the best F-measure in the baseline for the VIS dataset is

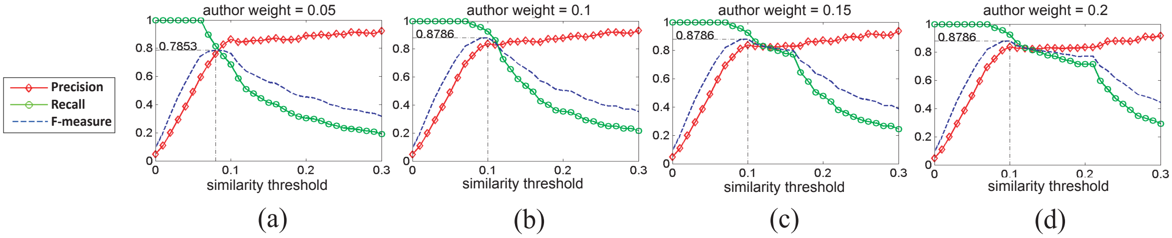


Fig. 3. Precision-recall of the VIS dataset: (a) w_m(author weight)$= 0.05$; (b) $w_m = 0.1$; (c) $w_m = 0.15$; (d) $w_m = 0.2$.

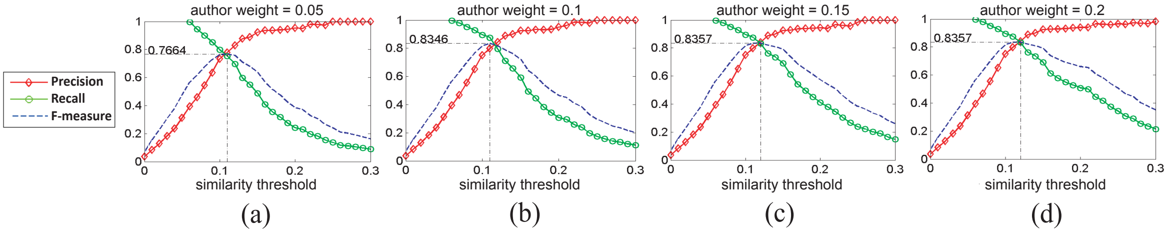


Fig. 4. Precision-recall of the DM dataset: (a) w_m(author weight)$= 0.05$; (b) $w_m = 0.1$; (c) $w_m = 0.15$; (d) $w_m = 0.2$.

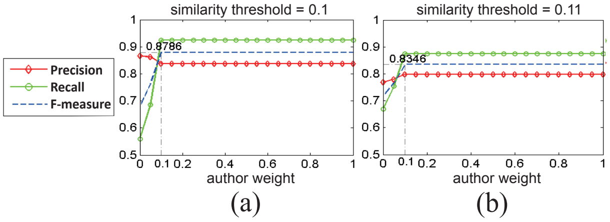


Fig. 5. Influence of the author weight: (a) VIS; (b) DM.

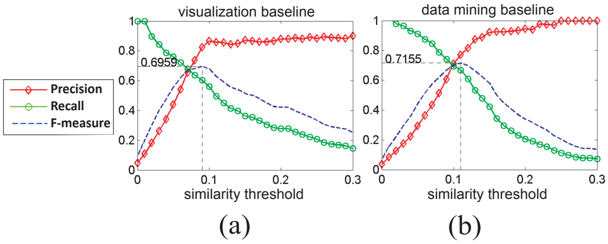


Fig. 6. Precision-recall of the baseline: (a) VIS; (b) DM.

0.6959, which is less than that of our approach (0.8786). The best F-measure in the baseline for the DM dataset is 0.7155, which is also less than that of our approach (0.8357). As a result, our approach achieved better results than the baseline.

5 Visualization

We developed TextPioneer through multiple sessions of participatory design with three professors and two NSF funding officers. In close collaboration with these experts at every stage of the visual design, we iteratively refined and improved the visual design and related visualization components. In this section, we introduce the components in detail.

5.1 Design Rationale

Our visualization design was grounded by interviewing five experts. Two of the experts manage funding portfolios in a research funding agency and the others are three professors in visualization, data mining, and social media analysis,respectively. The interviews were semi-structured with focuses on the participants' analysis processes and needs. Based on the interviews, we identified a two-phase analysis process that analyzes lead-lag at both global and local levels. Accordingly, we distilled the design requirements into global and local levels:

R1 (global) - scalable representation: One major concern of the experts is visual scalability since they often need to analyze two or three large corpora simultaneously. For large corpora with many topics, it is preferred to use a hierarchy to organize them. Accordingly, the experts emphasized the need to represent and explore the topic hierarchies.

R2 (global) - overall and individual corpus: All the experts hoped they could associate high-level corpora analysis with individual corpus examination. With this association, they could more easily find topics of interest for further exploration.

R3 (global, local) - lead-lag relationships: After understanding the global relationship, many tasks require further examination of the temporal lead-lag changes between two corpora in the context of specific topics.

R4 (global, local) - time and content comparison: The experts considered the time and the content of topics as important contextual information for understanding the lead-lag relationships at both global and local levels.

R5 (global, local) - context exploration: The experts emphasized the ability to explore the related contextual information, such as original documents and their metadata, on-demand for detailed analysis and hypothesis verification. In particular, they wanted to quickly identify the most influential documents/authors and inspect their influence, which helped the experts understand the major causes of the lead-lag relationship changes.

Based on the above requirements, we have designed the visualization components and related pipeline of TextPioneer (Fig. 7) . In the following sections, we present the design of each component in detail.

5.2 Global Lead-Lag as Hybrid Tree

It is a challenging task to design a visualization that meets all the requirements of global analysis. Before coming

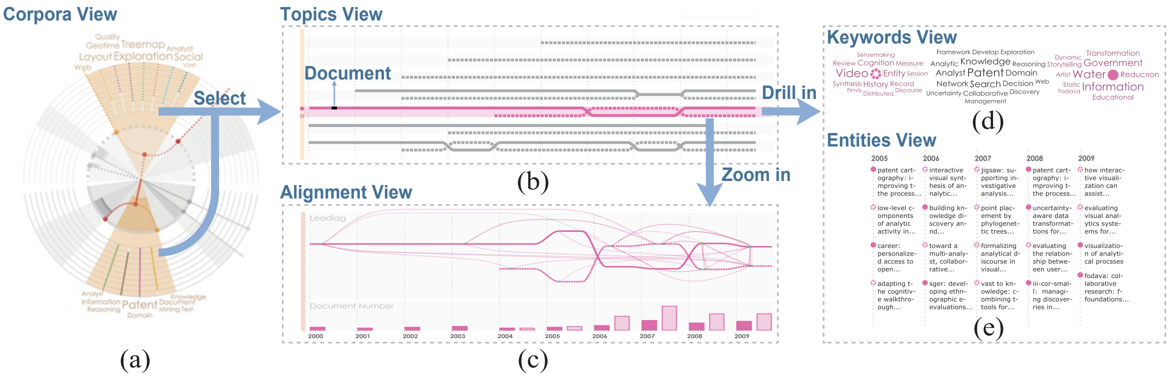


Fig. 7. Visualization pipeline: (a) explore the global lead-lag in the *corpora view*; (b) select the topics of interest and analyze their local lead-lag relationships in the *topic view*; (c) zoom in to the *alignment view* to examine the document alignment; (d) and (e) the *keywords* and *entities* views are synchronized with the *topic view*.

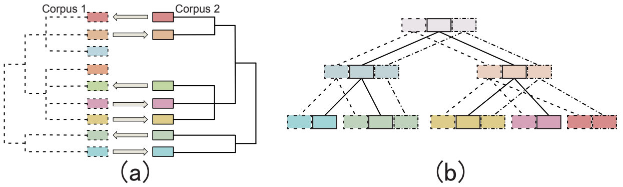


Fig. 8. Alternative designs for the *corpora view*: (a) side-by-side; (b) align three hierarchies by spatial proximity.

up with the current design, we had several failed trials. For example, we considered a straightforward design that places the hierarchies of two corpora side by side and uses directed links to convey their lead-lag relationships (Fig. 8(a)). However, such a side-by-side comparison is difficult to scale up for more than two corpora (R1). As an improvement, we aligned multiple hierarchies and placed their common nodes in a spatial proximity with decreasing lead values from left to right (Fig. 8(b)). Although the design can display multiple corpora, it lacked a clear picture for the individual corpus hierarchy (R2). As the data became larger, the layout also displays too many nodes and links, making the lead-lag exploration difficult (R3). Previous studies [29], [30] have shown that the radial space-filling tree visualization (Fig. 9(a)) is effective for exploring large, individual hierarchy (R2). Moreover, it utilizes the space efficiently, enabling the symmetrical visualization of multiple corpora within a single circular region (R1). As a result, we considered the radial space-filling tree as a potential solution. However, when displaying complex hierarchies, the traditional space-filling layout might hinder users from effective lead-lag comparison since the nodes to be compared are placed far away from each other due to their being multiple levels away from the root node (R3). For this reason, we propose a new design that combines a radial space-filling visualization with a node-link diagram (Fig. 9(b)). The basic idea is to display the focus nodes (L12, L21, and L22 in the upper corpus) and their children in the radial space-filling tree visualization while representing their ancestor nodes (non-focus nodes) in the node-link diagram. This combination ensures that the nodes and their counterparts being compared are placed more closely together.

Accordingly, we have developed *corpora view*, a hybrid tree visualization for visually comparing global lead-lag relationships across corpora (Fig. 9(c)). It consists of two major parts: an outer region that shows the topics being compared in a radial space-filling tree, along with an inner region that conveys the global lead-lag relationships and topic hierarchies (Fig. 10).

Outer region. In the outer region, multiple corpora are placed on the circumference, each of which is displayed within a sector called a **corpus sector**. In Fig. 10(a), two corpora are displayed and are encoded with solid and dotted

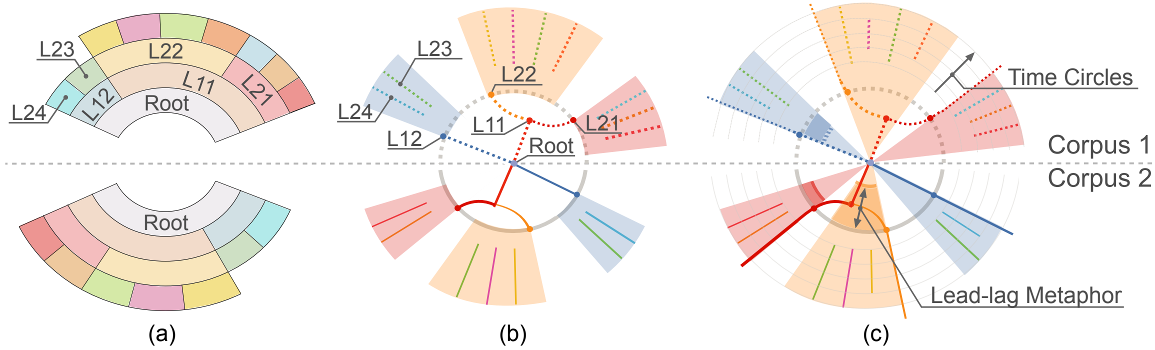


Fig. 9. The design of the *corpora view* (the same node is marked by the same annotation): (a) radial space-filling tree; (b) combine the radial space-filling tree (encoding the focus nodes such as L12, L21, and L22 in the upper corpus) with the node-link diagram (encoding the non-focus nodes); (c) a hybrid tree is generated by connecting the same nodes (from different corpora) along the same diagonal lines and overlaying the lead-lag metaphor.

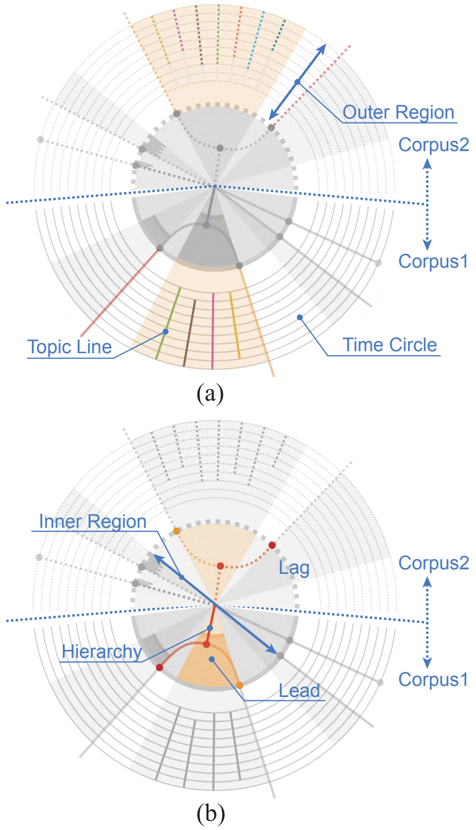


Fig. 10. Corpora view: (a) outer region; (b) inner region.

lines, respectively. Each corpus sector consists of multiple smaller sectors called **cluster sectors**. Each cluster sector displays a topic cluster. For example, in Fig. 10(a), four clusters are presented in the bottom corpus. The children of the cluster are displayed as **topic lines**. The outer region also contains a series of **time circles**, each of which represents a time unit, such as year. Each topic line starts from a time circle representing its starting year and ends at a time circle representing its ending year. A word cloud outside each cluster sector reveals its content. It consists of the most important keywords in each cluster (R4).

Color and position. The sectors with the same background color in different corpora contain shared topics. Shared topics are represented by topic lines of the same color. In each sector, shared topics are placed clock-wise in the same order, followed by unique topics in each corpus. The same layout criterion is also used to arrange the cluster sectors.

Inner region. The inner region represents the higher levels of the topic hierarchies using a node-link diagram, each node representing a cluster. It also conveys the global lead-lag relationships. The **lead** of the correlated clusters is encoded by the filled portions, while the **lag** is not (Fig. 10(b)). Larger filled portions indicate more significant leads.

5.3 Local Lead-Lag as Twisted Ladder

The lead-lag relationship between two corpora (regarding a specific topic) may change over time. Considering the characteristics of such change, we map the local lead-lag to the shape of a twisted ladder (R3). In the twisted-ladder-like visualization, the lead/lag of a topic are encoded by the spatial positions of the ladder rails since the spatial encoding is more quickly and accurately perceived than other encodings, such as color or size [31], [32]. Fig. 7(b) shows an example of the twisted-ladder-like visualization. A pair of twisted lines represents a shared topic in two corpora selected from the *corpora view*. The twisted line pairs are horizontally placed along the time dimension. The one on the top is the lead, while the one at the bottom is the lag. The twisted point indicates the lead and lag have been switched. A unique topic in a corpus is displayed as a single straight line.

To further examine documents that mainly cause the lead-lag change (R5), users can select a topic pair from the *topics view* and drill down to the *alignment view* to explore document alignment in the selected topic pair. As shown in Fig. 7(c), the vertical distance between the lines at each time encodes how significant the lead is. A document is represented by a dot on its topic line, whose horizontal position represents its time stamp. A curve connecting two dots indicates the content of the two documents is very similar (measured by the similarity score in Eq. (1)).

In addition to providing some basic interactions such as search, filtering, and selection, the *topics view* also allows users to iteratively refine a topic. The effectiveness of *TextPioneer* in supporting topic lead-lag analysis heavily depends on the quality of topic extraction. However, solely relying on topic modeling results is not effective since the automatic analysis is not always perfect and different users may have different requirements. To tackle this problem, *TextPioneer* allows users to refine the topic model by moving documents between topics. Specifically, when a document is selected, it is represented as a black dot on the corresponding topic, as shown in Fig. 7(b). Users can drag the dot to any topic line to move the document to that topic. Once the document is moved, the system performs lead-lag analysis on the new topic model in real time. The visual representation is also updated accordingly.

5.4 Coupling Topic with Entity and Word Content

To facilitate content and context exploration (R4, R5), *TextPioneer* provides two additional visual components: *entities view* (Fig. 7(e)) and *keywords view* (Fig. 7(d)). The *entities view* leverages a timeline to organize different types of the most influential entities, such as authors and documents, over time. Different metrics are used to sort different entities. For example, authors can be sorted by the total number of publications and/or citations. Based on the well-accepted word clouds layout, *keywords view* is used to convey the common and unique content of the topics of interest. For each topic, its keywords are classified into two categories: unique keywords that occur frequently in that topic but seldom in others, and common keywords that are owned by at least two selected topics. As shown in Fig. 7(d), the common keywords are displayed in black, while the unique keywords are encoded by the topic color that the keywords belong to.

6 LAYOUT ALGORITHM

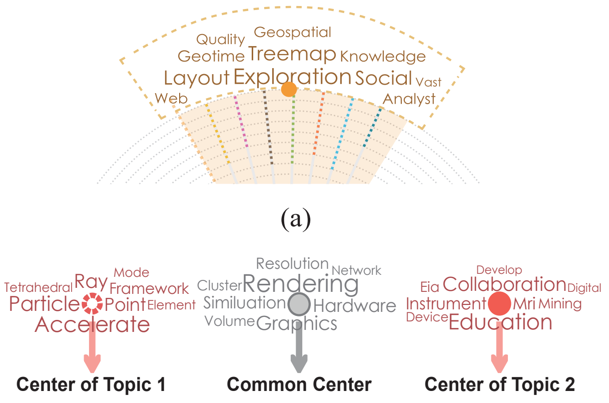
6.1 Hybrid Tree Layout

The layout of the hybrid tree can be simplified into a radial space-filling tree layout and a node-link diagram layout inside a circle. To enable users to see the topics of interest presented in full detail while at the same time get an overview of the tree, we adopt a focus + context technique. It first computes a Degree-Of-Interest (DOI) [33] for each node to measure its importance. When a node is selected, a maximum DOI is assigned. The DOI of other nodes decrease linearly with the distance from the node of the highest interest. Next, the nodes are displayed or elided according to the computed DOI. Their parent nodes are represented in the inner region using a node-link diagram. The sector area of a node is determined by its DOI. In case all the sector areas exceed the available space, we reduce the area of sectors with lower DOI values. These sectors can be shrunk, collapsed, and aggregated into their parent nodes, according to different DOI values.

6.2 Word Cloud Layout

In the word cloud, we aim to tightly pack as many keywords as possible to provide sufficient information for the users. On the other hand, we also want to maintain semantic similarity between keywords, as well as between keywords and topics. To this end, the layout method is based on a greedy sweep line algorithm [18]. It first finds a suitable space. Then the layout center is computed. Next, it generates a word cloud in the allowed area based on the center.

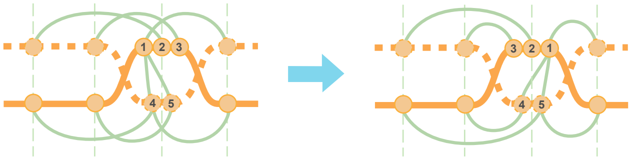
Finding the allowed space and layout center is the key to shaping a word cloud. In the *corpora view*, we use the fan area outside the topics as the layout area. To place the keywords closely to the topics, we define the center as the central point of the inner arc (Fig. 11(a)). In the *keywords view*, we treat the view area as the layout area. The major goal of the keyword view is to allow the user to examine the unique content (in keywords) of each selected topic, as well as the common content among them. Here, the unique content of a topic is the unique keywords in this topic, while the common content of the topics is the shared keywords among them. To this end, the layout method first builds a graph according to the content similarity of the selected topics. The content similarity is measured by the cosine similarity between keyword vectors of the topics. If the cosine similarity is greater than a given threshold, there will be an edge to connect the related topics. In our implementation, each topic is also connected to the common content in the topic graph. Next, the graph is laid out using a force-directed model [34]. Accordingly, the center of the unique keywords is its topic position determined by the force-directed layout, while the center of the common keywords is the center of the view (Fig. 11(b)).



(a)

(b)

Fig. 11. Word cloud layout: (a) *corpora view*; (b) *keywords view*.



(a)

(b)

Fig. 12. Ordering in document alignment: (a) random ordering; (b) ordering by average time stamps.

6.3 Layout of Document Alignment

To illustrate the content correlations between documents in the same topic across corpora, we model the documents and their relationships as nodes and links. Documents at each time are encoded by dots and are placed in its time range on each topic line, while links are represented by curves (Fig. 12(a)). The layout contains two steps:

Ordering: The ordering step is to determine the order of the documents at each time to minimize edge crossings. At each time, if we placed the documents that are aligned with later occurring documents in another corpus on the first position of the topic line, some unnecessary edge crossings will be introduced (Fig. 12(a)). To avoid this, we sort the document nodes at each time point by their average time stamps (Fig. 12(b)) and place them sequentially. The average time stamp of a document is calculated by averaging the time stamps of the documents that are correlated to it in another corpus. With this sorting strategy, edge crossings are minimized.

Merging: In our implementation, links with start and end points within a given distance (e.g., two-pixel) are merged to reduce visual clutter and accelerate rendering.

7 CASE STUDIES

We have applied *TextPioneer* to several text corpora, ranging from proposals/papers to social media. These applications aim to help experts like funding officers, researchers, business analysts, and sociologists to better understand the lead-lag patterns in different document collections. Here, two key applications are highlighted. The first application is for assessing the academic impact of scientific investments by a government agency (such as NSF or NASA), which we briefly looked at in the introduction. The second application aims to compare different social media such as blogs, message boards, and news to study information diffusion among multiple social communities.

7.1 NSF Proposal vs. Academic Publication

In this case study, domain experts explored a publication corpus and a proposal corpus to identify their lead-lag relationships. This study is used to illustrate how *TextPioneer* can be used by analysts to fulfill their analytical needs and point out what functions are useful for performing related tasks.

The publication corpus contains 10,125 paper abstracts collected from conference proceedings of Vis, InfoVis, VAST, KDD, NIPS, ICML, AAAI, SIGMOD, SIGIR, SIGGRAPH, and SIGGRAPH-ASIA from 2000 to 2009. The proposal corpus contains 1968 NSF proposal abstracts. They were extracted using keyword search from NSF grant proposals which are from 2000 to 2009. The keywords used were the most frequent keywords from the paper corpus. 76 topics were extracted for each corpus. They were organized into a 4-level hierarchy.

Two domain experts, who are computer science faculty, participated in the case study. They have been conducting research on visualization for 21 years and data mining for 15 years respectively. They have managed/co-managed projects funded by NSF and also published papers in the top conferences of their fields. Each of them spent one hour exploring the corpora with the instructors.

7.1.1 Investigating the Area of Visualization

The visualization expert examined the visualization related topics in the two corpora, as shown in Fig. 1(a). He recognized that there are two sub-areas of visualization. One is Scientific Visualization (SciVis) exemplified by keywords such as "Surface," "Volumetric," and "Isosurface." The other is Information Visualization (InfoVis) exemplified by keywords such as "Information," "Treemap," and "Layout." Overall, both sub-areas were led by proposals.

The expert explored the topics in the InfoVis sub-area in the *topics view*. A topic with keywords "Analytic," "Knowledge," and "Reasoning" attracted his attention. He commented that this is "Visual Analytics," the sub-area he co-founded. He opened this topic in the *alignment view* (Fig. 13). The histogram showed that before 2006, there were only a few NSF grants and publications on this topic. Since 2006, there was a burst of publications and they led the NSF grants. The expert indicated that this is due to the following three facts: 1) the research agenda of Visual Analytics was formed in 2005; 2) the first conference on Visual Analytics, the VAST conference, started in 2006; and 3) DHS was the primary funding source of this research field in the first several years after its research agenda came out. The *entities view* confirms that a large number of papers in this area were published in VAST. The *alignment view* also showed that after 2008, NSF grants began to lead publications. From the *entities view*, the expert discovered an NSF program named "Foundations Visual Analytics" in 2008 and 2009. He commented that the program beginning in 2008 promoted visual analytics research.

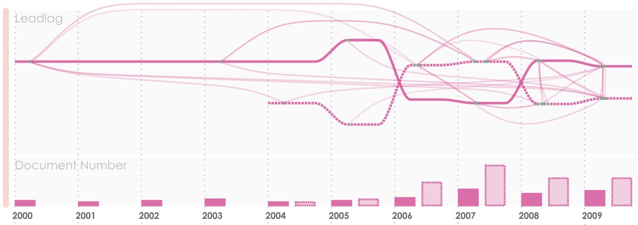


Fig. 13. The topical lead-lag evolution between proposals and publications. The topic is exemplified by keywords "Analytic/Knowledge/Reasoning."

7.1.2 Exploring the Area of Data Mining

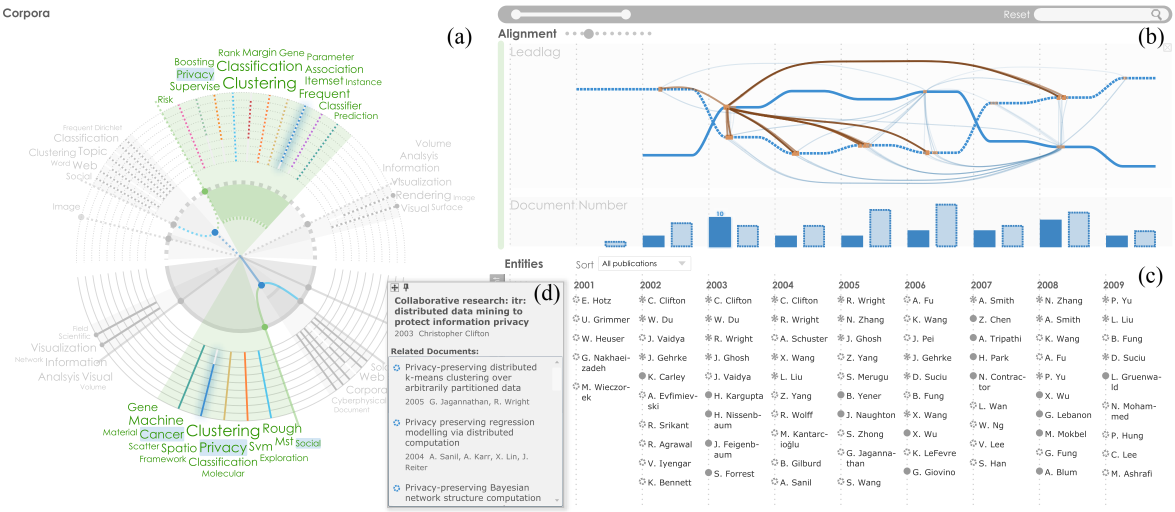
The data mining expert started by exploring a subarea of data mining, which is exemplified by "Clustering/Classification/Machine" (Fig. 14(a)). He noticed a topic exemplified by the keywords "Privacy" and "Social" (Fig. 14(a)), which concerns privacy preserving data mining. Since this was his recent research focus, he explored it in the *alignment view* for more details. At first glance, he noticed that academic publications led grant proposals before 2003 (Fig. 14(b)). He commented that this pattern is reasonable since privacy preserving data mining started from two well-known papers published at the Crypto and SIGMOD conferences in 2000 (the *alignment view* was blank in 2000 since we did not include papers from these two conferences). These papers were mainly authored by researchers from industry research labs so there was no NSF funding behind them. Then researchers in academia caught up. Papers such as "Privacy preserving mining of association rules" by J. Gehrke and "Privacy preserving association rule mining in vertically partitioned data" by C. Clifton were identified in 2002 from the *entities view*. In 2003, proposals influenced by these papers started to be funded by NSF and NSF grants began to lead publications. With the strong support from NSF, this research field bloomed.

In 2007, publications started to lead NSF grants again. The expert explained that J. Kleinberg gave the keynote speech "Challenges in mining social network data: processes, privacy, and paradoxes" at KDD 2007. This talk invoked new interest in this area and led to many novel papers. Detailed examination of the *entities view* supported his argument. For example, a paper entitled "Audience selection for on-line brand advertising: privacy-friendly social network targeting" was published in 2009.

The expert then examined authors from the *entities view*, as shown in Fig. 14(c). He immediately recognized several famous researchers in this field from the top of the lists, such as C. Clifton and J. Gehrke. The expert selected a grant proposal awarded to C. Clifton in 2003 (Fig. 14(d)). As shown in Fig. 14(b), the highlighted links indicate that it was influenced by a paper in 2002 and had significant impact on many papers published later.

7.1.3 Initial Feedback from NSF Program Managers

We also invited an NSF program manager and a former NSF program manager to evaluate *TextPioneer*. Both of



Entities Sort All publications

Collaborative research: itr: distributed data mining to protect information privacy
2003 Christopher Clifton

Related Documents:

Privacy-preserving distributed k-means clustering over arbitrarily partitioned data
2005 G. Jagannathan, R. Wright

Privacy preserving regression modelling via distributed computation
2004 A. Sanil, A. Karr, X. Lin, J. Reiter

Privacy-preserving Bayesian network structure computation

2001	2002	2003	2004	2005	2006	2007	2008	2009
E. Hotz	C. Clifton	C. Clifton	C. Clifton	R. Wright	A. Fu	A. Smith	N. Zhang	P. Yu
U. Grimmer	W. Du	W. Du	R. Wright	N. Zhang	K. Wang	Z. Chen	A. Smith	L. Liu
W. Heuser	J. Vaidya	R. Wright	A. Schuster	J. Ghosh	J. Pei	A. Tripathi	K. Wang	B. Fung
G. Nakhaei-zadeh	J. Gehrke	J. Ghosh	X. Wang	Z. Yang	J. Gehrke	H. Park	A. Fu	D. Suciu
M. Wieczor-ek	K. Carley	J. Vaidya	L. Liu	S. Merugu	D. Suciu	N. Contrac-tor	P. Yu	L. Gruenwa-ld
	A. Evfimiev-ski	H. Kargupta	Z. Yang	B. Yener	B. Fung	L. Wan	X. Wu	N. Moham-med
	R. Srikant	H. Nissenb-aum	R. Wolff	J. Naughton	X. Wang	W. Ng	G. Lebanon	P. Hung
	R. Agrawal	J. Feigenb-aum	M. Kantarc-ioğlu	S. Zhong	X. Wu	V. Lee	M. Mokbel	C. Lee
	V. Iyengar	S. Forrest	B. Gilburd	G. Jaganna-than	K. LeFevre	S. Han	G. Fung	M. Ashrafi
	K. Bennett		A. Sanil	S. Wang	G. Giovino		A. Blum	

Fig. 14. The lead-lag relationships between NSF grant proposals and academic publications in data mining: (a) *corpora view* shows publications lead proposals in the selected sub-area. A topic exemplified by keyword "Privacy" is highlighted; (b) *alignment view* shows the lead-lag evolution of the highlighted topic; (c) *entities view* shows the author information of the topic; (d) *document snippet* enables the detailed examination of the document.

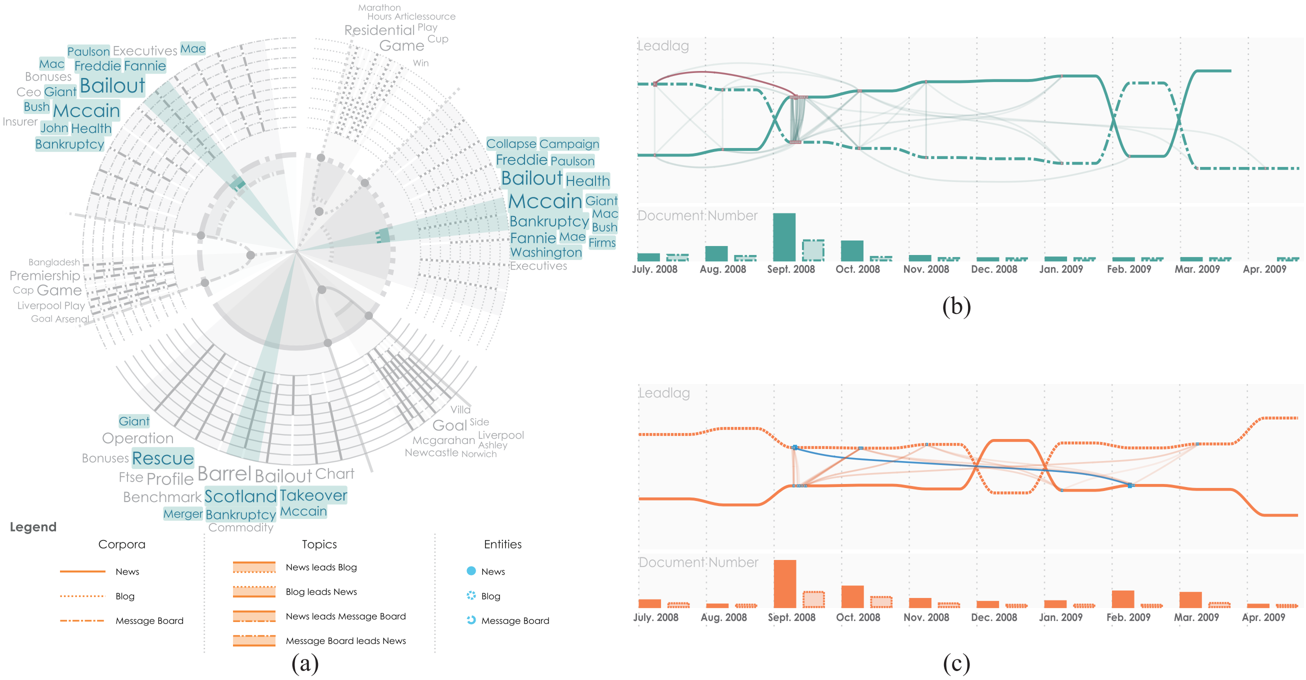


Fig. 15. The lead-lag relationships among news, blogs, and message boards; (a) *corpora view* shows the global lead-lag relationships. The blogs lead on most topics; (b) *alignment view* shows topical lead-lag evolution between message boards and news. The topic is exemplified by "Lehman/Market/Barclays;" (c) *alignment view* shows topical lead-lag evolution between blogs and news. The topic is exemplified by "AIG/Insurance/Billion."

them have more than two years of working experience at NSF. We conducted a semi-structured interview guided by a set of usability and effectiveness related questions. Each of the evaluations took 60 minutes, including 10 minutes of system demonstration, 30 minutes of case study and free exploration, and 20 minutes for the post interview.

The overall comments from the program managers were promising. They felt that *TextPioneer* was "a very useful exploratory tool for the NSF managers who manage new funding archives and interdisciplinary research programs." Program manager **A** pointed out "managing such unfamiliar research areas is always a great challenge and *TextPioneer* could help identify the influential researchers, important work, and the lead-lag relationships among the topics of interest." Program manager **B** said "It would be a great way to analyze all the NSF proposals (granted and non-granted ones) with this tool, especially those in the cross-edge research areas."

They also provided a lot of encouraging comments on the individual functions of *TextPioneer*. For example, both managers commented that the *alignment view* was extremely useful due to the ability to find previous influential publications for a proposal. The function can help NSF panelists further evaluate the novelty of the proposal, which is difficult to judge traditionally, especially for some interdisciplinary and emerging research areas. Manager **A** also liked the author/PI related information provided by the system. She said, "For me, they are more reliable than solely relying on the keyword computation." She suggested that the view could also help NSF managers identify leading researchers in a research area for forming review panels.

Furthermore, the managers suggested a few potential areas that could be improved. For example, they would like to examine the lead-lag relationships across different research areas and NSF proposals (e.g, granted and non-granted ones). They would also like to have more options for determining the lead influence of the document, such as evaluating them by citation networks or user specified keywords.

7.2 News vs. Blogs vs. Message Boards

In this case study, we aimed to demonstrate the scalability and effectiveness of *TextPioneer* by exploring lead-lag relationships among three large social media corpora, namely a news corpus, a blog corpus, and a message board corpus. The news corpus contained 66,370 articles. The blog corpus consisted of 16,520 blogs. The message board corpus contained 34,869 posts. All the documents were collected from Boardreader (http://boardreader.com/) using twenty financial companies' names, such as "AIG insurance," "Barclays," and "Bank of America" (from Jul. 2008 to Apr. 2009). 22 topics, were identified by *TextPioneer*. The news corpus was set as the reference corpus.

As shown in Fig. 15(a)), the blogs led message boards and news on most topics globally. For the other two corpora, message boards led news on breaking events such as "Lehman/Market/Barclays" and "AIG/Insurance/Billion," while lagging behind on sports topics such as "Game/Premier/Team" and "Barclays/League/Club."

We then investigated local lead-lag changes for the topics of interest. The *alignment view* in Fig. 15(b) shows the changes between the message boards and the news on the topic "Lehman/Market/Billion". Message boards led the news until Sept. 2008, when Lehman Brothers declared bankruptcy. The highlighted link indicates that the message boards had started to discuss the rumors of Lehman Brothers' bankruptcy, such as one that indicated Lehman Brothers would be sold to Barclays, another big bank, as early as Jul. 2008. After Sept. 2008, a great deal of news on the financial crisis and bank bankruptcy was reported, leading to in-depth discussions on message boards. For example, people discussed whether the bank bankruptcy would spread to UK banks in Oct. 2008.

Fig. 15(c) shows the lead-lag changes between blogs news on topic "AIG/Insurance/Government". The histogram shows a burst of news in Sept. 2008. Examining the news articles from the *entities view* shows that they are about the bankruptcy of the insurance company AIG and the government's bailout of AIG. Then people began to discuss the illegal dealings and scams of AIG on blogs. By tracking the highlighted link, we found that the news did not respond to these issues until Feb. 2009.

To evaluate the usefulness of our approach on social media data, we conducted a semi-structured interview with two experts, a sociology PhD student (User S) and a professor (User P) in media and communication studies.

Overall, *TextPioneer* has been well received by the experts. They feel that the toolkit is very useful for examining lead-lag relationships among corpora. User P commented, "The Corpora view is very important for me to get an overview. After knowing the overall patterns, what I am interested in are the finer-grained lead-lag relationships at the topic level. The hierarchies in the Corpora view and the twisted-ladder-like visualization in the Topics view enable me to find the detailed topics and their lead-lag changes quickly. This is a great way for me to find something unexpected." The experts were further impressed by the local lead-lag examination function provided by *TextPioneer*. For example, User S was intrigued when he found the topic pair from different corpora changed their lead-lag relationships at several time points. He said, " Social media sites such as Twitter are reshaping the landscape of journalism. More and more breaking news come from social media first. I believe this system can definitely help a journalist find the best social media outlet to collect breaking news." User S was also impressed by the visualization and commented, "It is engaging and useful, and definitely outperforms the static infographics that I used before."

The experts also suggested several potential improvements for *TextPioneer*. For example, User P suggested adding sentiment information to the twisted-ladder-like visualization. With this, users can easily examine the public opinion at different time points and compare the opinion before and after the lead changes.

8 Discussion

In case studies and interviews with experts, we collected a great deal of feedback. The positive feedback is summarized according to the following areas:

Overall visualization design: All the experts were motivated by the interactive visualization of *TextPioneer* when exploring the lead-lag relationships across corpora. The corpora view at first seemed a little complex to some of them (two professors and one NSF officer). However, after engaging with it for 5-10 minutes, the experts thought the visualization was appealing and inviting because it helps them easily find lead-lag patterns. All the experts liked the design of the hybrid visualization, which allows them to easily examine the topics at different levels. One of them said, "I like such a compact circular layout, which helps me easily compare the topic content and the temporal information of topics. This information is really essential to understand and verify the [lead-lag] results and trigger further exploration." In the *corpora view*, they particularly liked the hierarchical representation of the topics and one of them stated, "It's great to have the hierarchy, especially for the corpus with huge amounts of text documents. This is exactly what I would expect for a text visualization tool."

According to the interview of the experts, they can leverage TextPioneer to handle two and three corpora very well. The *corpora view* also works for four corpora although it takes longer to gain insight. It may fail to provide a better understanding for five or more corpora due to the limited display area. Although the *topic view* can only display alignments between the shared topics in two corpora, it is good enough to support the analysis tasks of our target users at the detailed level. Previous experiments [35], [36] have consistently found that approximately four objects can be tracked in visual comparison. As a result, TextPioneer works for most real-world applications.

Regarding the local analysis, all the experts thought the twisted-ladder-like visualization provided "very intuitive visual encodings" for inspecting the lead-lag changes and enabled a "concise and clear representation" by hiding the details. The word cloud allowed them to "compare the topic content quickly and effectively." Moreover, two experts found the ability to sort the entities incredibly useful since it enabled them to "immediately find some leading researchers or important papers that influence the development of the research field."

Capability of insight discovery: Overall, all the experts were able to easily use *TextPioneer* to perform detailed lead-lag exploration and derive many interesting insights. They were impressed by the power of the visualization components in identifying lead-lag patterns across corpora. For example, one of the professors expressed a strong interest in using *TextPioneer* in his own research for analyzing large social media data. He believed *TextPioneer* can help him derive interesting relationships among different social media outlets, which is beyond the functionality of the visualization tools he currently employs. Another professor commented, "The first time I saw the [alignment] view, I was attracted by the compelling interface. It encourages me to go through [the view] and match the patterns to my mental picture of this research field. Then I realized that I can smoothly interact with the tool and table, which helps me verify the patterns and understand the underlying analysis model. This is exactly the value of the system."

Besides the above positive feedback, the experts also indicated some limitations of *TextPioneer*. First, a lot of information is compacted into the *corpora view*. This is caused by the complexity of the lead-lag analysis results and task requirements. If we provide too little information, the analyst may have little to go on. On the other hand, if we provide too much information, the visualization is cluttered and hardly anything can be inferred from it. To solve this problem, we use preattentive visual channels to encode the important data attributes and use other channels to represent the less important ones. In addition, we only display the important information at first. Other information will be displayed on demand. Second, it is not easy to compare the lead-lag relationships across different research topics within one corpus, which is very useful in interdisciplinary research. One straightforward method is to treat the two or more selected topics as the same topic across different corpora and then apply the lead-lag analysis method. Another systematic solution is to model the influence among different topics. This is a good point for future investigation and we mention it in Sec. 9.

The experts also suggested several potential applications of *TextPioneer*, including product/brand promotion by tracking the related opinion diffusion across different social outlets, opinion leader identification in social media, and discovery of important information or breaking news stories for news reports.

9 Conclusions and Future Work

In this paper, we present an interactive, visual lead-lag analytics system, called *TextPioneer*. The major feature of this tool is that it allows users to visually analyze topical lead-lag relationships both globally and locally. *TextPioneer* provides three significant benefits over previous methods. First, it derives topical lead-lag relationships from global patterns to local changes over time. Second, it provides a consistent visualization to intuitively illustrate two-level results. Third, *TextPioneer* offers users a set of rich interactions that enable them to reason about lead-lag analysis results in context. Through two case studies, we demonstrated the usability and usefulness of our system in tackling global and local lead-lag across corpora, especially in exploring the major factors that lead to such patterns.

We are also working on several areas to further improve *TextPioneer*. According to domain experts, the development of a topic is influenced by the same topic in other corpora and/or the related topics within the same corpus or across corpora. Thus, the first area of possible improvement is to analyze topic influence within each corpus, as well as across corpora. Second, we would like to study how to leverage various application-specific features to derive more accurate leads and lags, and their changes over time. Third, we plan to design a learning-based method to automatically estimate most of the weights in lead-lag analysis.

References

[1] R. Nallapati, X. Shi, D. A. McFarland, J. Leskovec, and D. Jurafsky, "Leadlag lda: estimating topic specific leads and lags of information outlets," in *ICWSM*, 2011, pp. 558–561.

[2] X. Shi, R. Nallapati, J. Lescovec, D. A. McFarland, and D. Jurafsky, "Who leads whom: topical lead-lag analysis across corpora," in *Neural Information Processing Systems Workshop on Computational Social Science and Wisdom of Crowds*, 2010.

[3] J. Zhang, Y. Song, C. Zhang, and S. Liu, "Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora," in *KDD*, 2010, pp. 1079–1088.

[4] L. Lloyd, P. Kaulgud, and S. Skiena, "Newspapers vs. blogs: who gets the scoop?" in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 117–124.

[5] J. Leskovec, L. Backstrom, and J. M. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *KDD*, 2009, pp. 497–506.

[6] T. Menezes, C. Roth, and J.-P. Cointet, "Precursors and laggards: an analysis of semantic temporal relationships on a blog network," in *SocialCom/PASSAT*, 2010, pp. 120–127.

[7] S. Gerrish and D. M. Blei, "A language-based approach to measuring scholarly impact," in *ICML*, 2010, pp. 375–382.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[9] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou, "Treejuxtaposer: scalable tree comparison using focus+context with guaranteed visibility," *ACM TOG*, vol. 22, no. 3, pp. 453–462, 2003.

[10] S. Bremm, T. von Landesberger, M. Hess, T. Schreck, P. Weil, and K. Hamacher, "Interactive visual comparison of multiple trees," in *IEEE VAST*, 2011, pp. 31–40.

[11] G. G. Robertson, M. Czerwinski, and J. E. Churchill, "Visualization of mappings between schemas," in *CHI*, 2005, pp. 431–439.

[12] D. Holten and J. J. van Wijk, "Visual comparison of hierarchically organized data," *Comput. Graph. Forum*, vol. 27, no. 3, pp. 759–766, 2008.

[13] M. Graham and J. B. Kennedy, "Exploring multiple trees through dag representations," *IEEE TVCG*, vol. 13, no. 6, pp. 1294–1301, 2007.

[14] Y. Tu and H.-W. Shen, "Visualizing changes of hierarchical data using treemaps," *IEEE TVCG*, vol. 13, no. 6, pp. 1286–1293, 2007.

[15] P. Isenberg and M. S. T. Carpendale, "Interactive tree comparison for co-located collaborative information visualization," *IEEE TVCG*, vol. 13, no. 6, pp. 1232–1239, 2007.

[16] S. Havre, E. G. Hetzler, P. Whitney, and L. T. Nowell, "Themeriver: visualizing thematic changes in large document collections," *IEEE TVCG*, vol. 8, no. 1, pp. 9–20, 2002.

[17] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian, "Interactive, topic-based visual text summarization and analysis," in *CIKM*, 2009, pp. 543–552.

[18] L. Shi, F. Wei, S. Liu, L. Liu, X. Lian, and M. X. Zhou, "Understanding text corpora with multiple facets," in *IEEE VAST*, 2010, pp. 99–106.

[19] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian, "Tiara: Interactive, topic-based visual text summarization and analysis," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 2, pp. 25:1–25:28, 2012.

[20] Z. Gao, Y. Song, S. Liu, H. Wang, H. Wei, Y. Chen, and W. Cui, "Tracking and connecting topics via incremental hierarchical dirichlet processes," in *ICDM*, 2011, pp. 1056–1061.

[21] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "Textflow: towards better understanding of evolving topics in text," *IEEE TVCG*, vol. 17, no. 12, pp. 2412–2421, 2011.

[22] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. A. Keim, "Eventriver: visually exploring text collections with temporal references," *IEEE TVCG*, vol. 18, no. 1, pp. 93–105, 2012.

[23] K. Wongsuphasawat and D. Gotz, "Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization," *IEEE TVCG*, vol. 18, no. 12, pp. 2659–2668, 2012.

[24] S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu, "Storyflow: Tracking the evolution of stories," *IEEE TVCG*, vol. 19, no. 12, pp. 2436–2445, 2013.

[25] Y. Tanahashi and K.-L. Ma, "Design considerations for optimizing storyline visualizations," *IEEE TVCG*, vol. 18, no. 12, pp. 2679–2688, 2012.

[26] Y. W. Teh, M. I. Jordan, M. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 476, no. 101, pp. 1566–1581, 2006.

[27] J. Zhang, Z. Ghahramani, and Y. Yang, "A probabilistic model for online document clustering with application to novelty detection," in *NIPS*, 2005, pp. 1617–1624.

[28] C. Blundell, Y. W. Teh, and K. A. Heller, "Bayesian rose trees," in *UAI*, 2010, pp. 65–72.

[29] J. T. Stasko, R. Catrambone, M. Guzdial, and K. Mcdonald, "An evaluation of space-filling information visualizations for depicting hierarchical structures," *Int. J. Hum.-Comput. Stud.*, vol. 53, no. 5, pp. 663–694, 2000.

[30] M. Burch, N. Konevtsova, J. Heinrich, M. Höferlin, and D. Weiskopf, "Evaluation of traditional, orthogonal, and radial tree diagrams by an eye tracking study," *IEEE TVCG*, vol. 17, no. 12, pp. 2440–2448, 2011.

[31] J. D. Mackinlay, "Automating the design of graphical presentations of relational information," *ACM TOG*, vol. 5, no. 2, pp. 110–141, 1986.

[32] P. McLachlan, T. Munzner, E. Koutsofios, and S. North, "Liverac: interactive visual exploration of system management time-series data," in *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ser. CHI '08. New York, NY, USA: ACM, 2008, pp. 1483–1492.

[33] G. W. Furnas, "Generalized fisheye views," in *CHI*, 1986, pp. 16–23.

[34] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs," *Inf. Process. Lett.*, vol. 31, no. 1, pp. 7–15, 1989.

[35] J. Intriligator and P. Cavanagh, "The spatial resolution of visual attention," *Cognitive psychology*, vol. 43, no. 3, pp. 171–216, 2001.

[36] S. Yantis, "Multielement visual tracking: Attention and perceptual organization," *Cognitive psychology*, vol. 24, no. 3, pp. 295–340, 1992.

Shixia Liu is a lead researcher in the Internet Graphics Group at Microsoft Research Asia. Her research interests include interactive, visual text analytics and interactive, visual social analytics. She received a B.S. and M.S. in Computational Mathematics from Harbin Institute of Technology, a Ph.D. in Computer Science from Tsinghua University. Before she joined MSRA, she worked as a research staff member at IBM China Research Lab.

Yang Chen is a PhD student in the Computer Science Department at the University of North Carolina at Charlotte. He received a BS degree in Computer Science from Wuhan University. His research interests include visual analytics on high dimensional data and text documents.

Hao Wei is a graduate student in the State Key Lab of CAD&CG at Zhejiang University, China. She received a BS degree in Computer Science from Zhejiang University. Her research interests include visual text analytics and text mining.

Jing Yang is an associate professor in the Computer Science Department at the University of North Carolina at Charlotte. Her research interests include visual analytics and information visualization on high dimensional data, graphs, and text documents. She received a PhD in computer science from Worcester Polytechnic Institute.

Kun Zhou is currently a Cheung Kong Distinguished Professor in the Computer Science Department of Zhejiang University, and a member of the State Key Lab of CAD&CG. Before joining Zhejiang University, he was a Lead Researcher of the graphics group at Microsoft Research Asia. He received his BS degree and PhD degree in computer science from Zhejiang University in 1997 and 2002, respectively. His research interests include shape modelling/editing, texture mapping/synthesis, real-time rendering and GPU parallel computing.

Steven M. Drucker is a Principal Researcher in the Visualization and Interaction group at Microsoft Research (MSR) focusing on human computer interaction for dealing with large amounts of information. He is also an affiliate professor at the University of Washington. Before coming to Microsoft, he received his Ph.D. from the Computer Graphics and Animation Group at the MIT Media Lab in May 1994 and a M.S from the AI Lab at MIT in 1989.