# TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis

SHIXIA LIU, Microsoft Research Asia
MICHELLE X. ZHOU and SHIMEI PAN, IBM Research
YANGQIU SONG, Microsoft Research Asia
WEIHONG QIAN, WEIJIA CAI, and XIAOXIAO LIAN, IBM Research

We are building an interactive visual text analysis tool that aids users in analyzing large collections of text. Unlike existing work in visual text analytics, which focuses either on developing sophisticated text analytic techniques or inventing novel text visualization metaphors, ours tightly integrates state-of-the-art text analytics with interactive visualization to maximize the value of both. In this article, we present our work from two aspects. We first introduce an enhanced, LDA-based topic analysis technique that automatically derives a set of topics to summarize a collection of documents and their content evolution over time. To help users understand the complex summarization results produced by our topic analysis technique, we then present the design and development of a time-based visualization of the results. Furthermore, we provide users with a set of rich interaction tools that help them further interpret the visualized results in context and examine the text collection from multiple perspectives. As a result, our work offers three unique contributions. First, we present an enhanced topic modeling technique to provide users with a time-sensitive and more meaningful text summary. Second, we develop an effective visual metaphor to transform abstract and often complex text summarization results into a comprehensible visual representation. Third, we offer users flexible visual interaction tools as alternatives to compensate for the deficiencies of current text summarization techniques. We have applied our work to a number of text corpora and our evaluation shows promise, especially in support of complex text analyses.

Categories and Subject Descriptors: H.5.2 [**User Interfaces**]: Graphical User Interfaces (GUI)

General Terms: Design, Human Factors

Additional Key Words and Phrases: Text analytics, interactive text visualization, stacked graph, text trend chart, text summarization, topic model

**ACM Reference Format:**
Liu, S., Zhou, M. X., Pan, S., Song, Y., Qian, W., Cai, W., and Lian, X. 2012. TIARA: Interactive, topic-based visual text summarization and analysis. ACM Trans. Intell. Syst. Technol. 3, 2, Article 25 (February 2012), 28 pages.
DOI = 10.1145/2089094.2089101 http://doi.acm.org/10.1145/2089094.2089101

## 1. INTRODUCTION

We live amidst seas of text documents, including academic publications, news articles, emails, and patient records. We use them to convey information, share knowledge, coordinate activities, as well as to record business conduct. In many lines of work, we



Authors' addresses: S. Liu (corresponding author), Microsoft Research Asia; email: shixia@gmail.com; M. X. Zhou and S. Pan, IBM Research; Y. Song, Microsoft Research Asia; W. Qian, W. Cai, and X. Lian, IBM Research.

DOI 10.1145/2089094.2089101 http://doi.acm.org/10.1145/2089094.2089101

are often required to swiftly analyze large collections of documents as part of our job. For example, healthcare givers must analyze seas of patient records to manage care resources; business auditors need to sift through mountains of documents (e.g., email archives) to ensure business compliance.

To help people cope with the ever-increasing amounts of text documents, researchers have developed advanced technologies facilitating text analyses. Such efforts are mainly from two research communities: text analytics and information visualization. From the text analytics community, researchers have developed a wide array of text analysis algorithms (e.g., Carenini et al. [2007], Dredze et al. [2008], Wan and McKeown [2004]). From the information visualization community, researchers have designed a number of text visualization techniques (e.g., Nardi et al. [2002], Perer and Smith [2006], Viegas et al. [2006]).

However, there are few efforts that tightly couple state-of-the-art text analytics with interactive visualization to maximize the value of both. For example, researchers in text analytics use only basic visualizations to display their final analysis results (e.g., matrix visualization in McCallum et al. [2007] and scatter plots in Iwata et al. [2008]). On the other hand, researchers in information visualization focus on illustrating rather simple analysis results (e.g., TFIDF measure of keywords in Viegas et al. [2006]).

Although these existing techniques have achieved a certain amount of success, they may be inadequate in support of many real-world text analysis tasks. Consider the role of a customer relationship manager of a hotel chain, who is examining customers' opinions about the chain. To do so, s/he must analyze a large collection of texts, customer feedback posted online or sent via emails, to answer a set of questions, including the following.

— What are the major topics in the customer feedback?
— What are the most active topics over the last three months?
— What are the key concepts mentioned in the aforesaid topics?
— How have the most active topics evolved over time?

To help users answer such questions, we tightly integrate interactive visualization with state-of-the-art text summarization techniques to visually summarize a large text corpus. Specifically, we are building an interactive visual text analysis tool, called TIARA (Text Insight via Automated, Responsive Analysis). Previously, we have described how TIARA can automatically generate a visual summary of text analytic results [Liu et al. 2009]. Here, in this article we focus on TIARA's visual text analytic lifecycle, especially on TIARA's enhanced topic modeling method and its support of interactive visual analysis of huge document corpora. The following three key aspects of TIARA are described in this article.

First, TIARA uses an enhanced topic modeling engine that aims to provide users with a meaningful, time-sensitive topic-based summary. Our engine first uses an LDA-based topic modeling technique [Blei et al. 2003] to summarize the documents into a set of topics, each of which is represented by a set of keywords. Given the derived topics, it then computes the top-N most salient topics for users to gain an overall understanding of a text collection. To depict the content evolution of each topic over time, our engine also derives the top-K most salient keywords to describe the topic at every given time point.

Second, TIARA uses a time-based visualization to explain text summarization results derived by its text analytic engine. The topic analysis output is complex, including a set of topics, each of which is depicted by a set of keywords and their

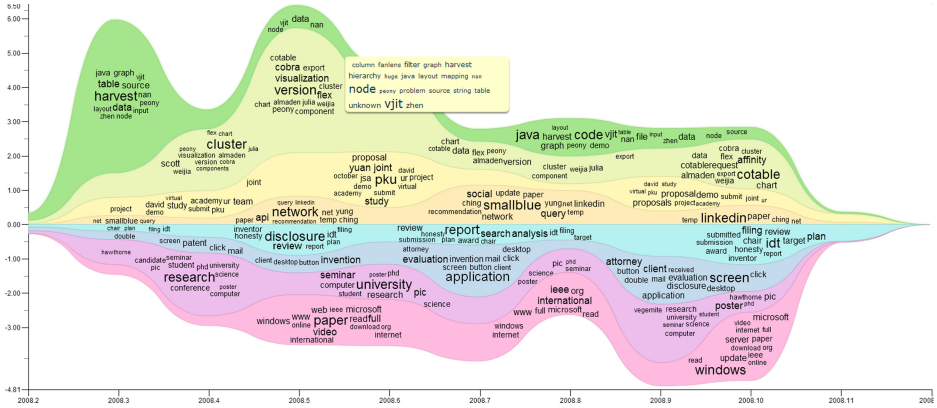


Fig. 1. TIARA-created visual summary of about 10,000 emails in 2008. The x-axis encodes time, and the y-axis encodes topic strength. Each layer represents a topic and its evolution over time through a set of keywords distributed along the time line. The tool tip shows the overall, aggregated keywords of the top-most topic (green one).

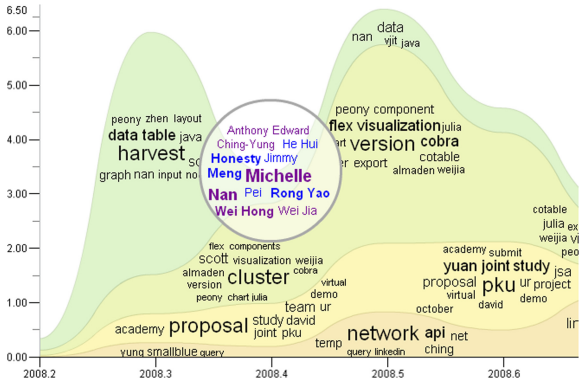


Fig. 2. A lens showing the names of email senders and receivers under two topics around April, 2008. Color is used to distinguish people from different organizations.

probabilistic distributions. To make the abstract and complex results consumable by average users, we design a time-based, topic-oriented visualization. Figure 1 shows such a visualization created by TIARA. Each colored layer represents a derived topic. Each layer is depicted by a set of keyword clouds, summarizing the topic content and the content evolution over time. The width of a layer at a time point encodes the "strength" of the topic and the number of documents covering the topic at that time.

Third, TIARA provides users with rich interaction tools that allow them to further interpret and examine summarized text from multiple perspectives. Today's text summarization techniques are less than perfect. The interaction tools provide users with alternatives to compensate for the deficiencies in these techniques. For example, a user may not understand a derived topic and its associated keywords. In such cases, the user could acquire additional information through a magic lens (Figure 2). Moreover, the user can investigate the meanings of a topic keyword in the context of relevant text messages (Figure 3).

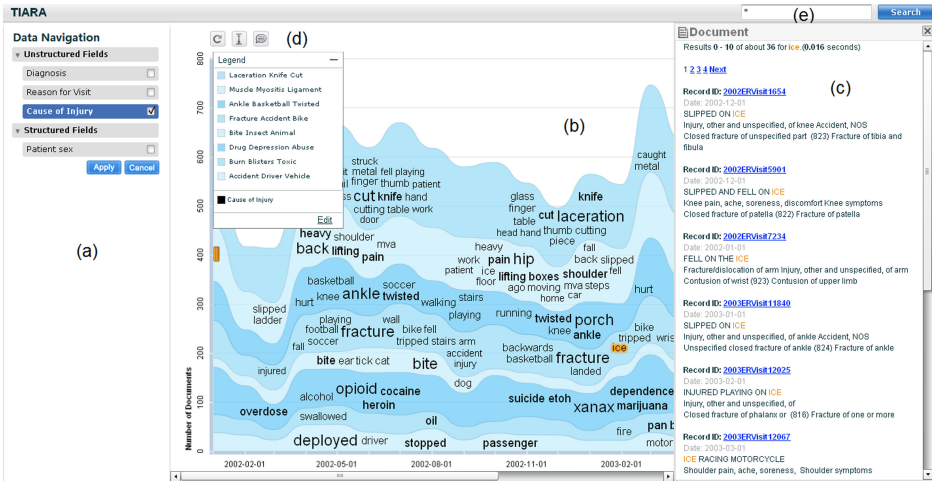


Fig. 3. The user interface of TIARA with a visual summary of the "reason for visit" field of the 23,000+ patient emergency room records from the the National Hospital Ambulatory Medical Care Survey dataset.

The goal of our work is to combine sophisticated, automatic text summarization techniques with interactive visualization to support an iterative, progressive text analysis. As a result, our work offers three unique contributions.

— *Enhanced topic modeling techniques* can produce time-sensitive and more meaningful text summaries.
— *Effective visual metaphors* allow users to comprehend abstract and complex text summaries and aid them in complex text analyses.
— *Flexible visual interaction techniques* let users interact with a visual text summary in context and examine texts from multiple aspects to compensate for the deficiencies of current text summarization technologies.

In the rest of the article, we first discuss related work. We then give an overview of TIARA before presenting TIARA's key features, followed by its example applications, and our evaluation.

## 2. RELATED WORK

Our work is directly related to the research efforts in two areas: text analytics and information visualization.

In the area of text analytics, researchers have developed a number of approaches to text summarization (e.g., Dredze et al. [2008], Carenini et al. [2007], McCallum et al. [2007], Wan and McKeown [2004], Wang et al. [2008]). There are two main techniques: sentence-based [Carenini et al. 2007; Wan and McKeown 2004; Wang et al. 2008] and keyword-based [Dredze et al. 2008; McCallum et al. 2007] text summarization. Sentenced-based approaches identify the most salient sentences in a document [Carenini et al. 2007; Wan and McKeown 2004; Wang et al. 2008]. However, it may be time consuming for users to read several sentences per document especially when handling a large number of documents. Alternatively, keyword-based methods summarize documents by topics, each of which is characterized by a set of keywords [Dredze et al. 2008; McCallum et al. 2007]. TIARA's text summarization is built on the latter method, but its focus is on enhancing the summarization results through topic/keyword ranking and visualization. Moreover, we provide users with visual interaction tools to examine the results from multiple perspectives.

In the area of information visualization, researchers have developed various visualization approaches to text analysis (e.g., Perer and Smith [2006] and Viegas et al. [2006]). Based on the type of information being visualized, these systems can be classified into two categories: metadata-based and content-based text visualization. Metadata-based text visualization focuses on visualizing the metadata of text documents. For example, in email analysis, there are thread-based email visualizations [Kerr 2003; Venolia and Neustaedter 2003], and relationship-based visualizations of email senders and receivers (e.g., Nardi et al. [2002] and Perer and Smith [2006], www.enronexplorer.com, and jheer.org/enron). Similarly, in text search, there is visualization of document metadata, including document length and query term frequency [Hearst 1995].

Although metadata-based visualization aids in text analysis, it is inadequate in uncovering deeper insights often buried in the text. Specifically, it does not work for documents with little metadata. Thus, researchers have developed content-based text visualization. For example, Viegas et al. use Themail to visualize keywords based on their TFIDF scores in an email collection [Viegas et al. 2006]. Similarly, Strobelt et al. use a mixture of images and TFIDF-based keywords to create a compact visualization of a document [Strobelt et al. 2009]. There are also many other general-purpose text visualization systems that transform a collection of text into a visual illustration [Wise et al. 1995]. These systems include Galaxy of News [Rennison 1994], Jigsaw [Stasko et al. 2008], and WordTree [Wattenberg and Viegas 2008].

Based on the information being visualized, there are two types of content-based text visualization. The first type, including Galaxy of News [Rennison 1994] and Jigsaw [Stasko et al. 2008], focuses on depicting document relationships. More recently, Chen et al. [2009] and Iwata et al. [2008] focus on visualizing document clustering results. In contrast, the second type illustrates text content at the word or phrase level, including TextArc (www.textarc.org), WordTree [Wattenberg and Viegas 2008], Phrase Net [van Ham et al. 2009], and FeatureLens [Don et al. 2007]. Another system, MemeTracker, is developed to track the changes of short phrases over time [Leskovec et al. 2009].

Similar to our work, ThemeRiver [Havre et al. 2002] visually depicts the thematic variations over time within a document collection. While ThemeRiver uses a river metaphor to illustrate only the thematic strength and focuses on the "river" aesthetics (e.g., visual symmetry), TIARA uses a stack graph to convey much richer and far more complex thematic content. To do so, not only does TIARA illustrate thematic strength variations over time, but it also depicts the detailed thematic content in keywords (Figure 1). Furthermore, TIARA is designed to visualize fully machine-derived, complex text summarization results, including the imperfect ones. As described later, this goal of ours also poses additional visualization challenges that TIARA must address.

Compared with existing text visualization systems, TIARA visualizes both text content and metadata to facilitate a more comprehensive text analysis. It tightly combines state-of-the-art text summarization techniques with interactive visualization to support a more in-depth and practical text analysis. On the one hand, TIARA uses visualizations to convey complex summarization results and make them comprehensible. On the other hand, it offers rich visual interaction tools for users to examine texts to compensate for the deficiencies of the summarization technology.

## 3. TIARA OVERVIEW

TIARA is designed to visually summarize the topics/themes existing in a document collection and their content changes over time to facilitate text analysis. Here we provide an overview of TIARA, starting with its user interface, followed by its system architecture.

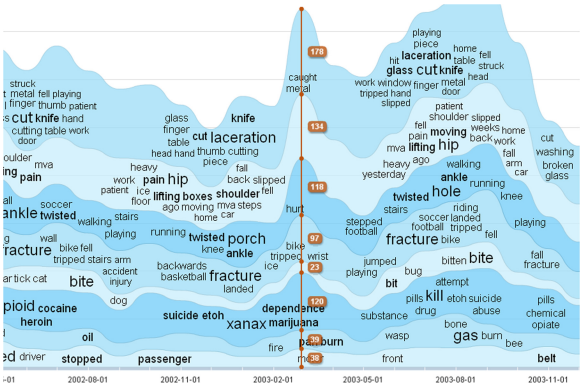


Fig. 4. Display of topic strengths at a selected time point.

### 3.1. User Interface

As shown in Figure 3, TIARA provides users with five main interaction areas: data navigation (Figure 3(a)), topic view (Figure 3(b)), document view (Figure 3(c)), action menu (Figure 3(d)), and search box (Figure 3(e)). When interacting with TIARA, a user may enter a query term in the search box (Figure 3(e)) to retrieve a document collection. The search results are presented in two views: the topic view (Figure 3(b)) and the document view (Figure 3(c)).

In an effort to reduce visual clutter in the display, the topic view (Figure 3(b)) initially shows only the top-N most salient topics derived from a document collection. The user can explicitly request any topics of interest using the interactive topic legend (Figure 3(d)). The topic view allows users to examine the topics from several perspectives, including viewing the thematic content of a topic shown in keywords at different time intervals (Figure 1), as well as all the thematic content of a topic in a tool tip (Figure 1).

Complementing the topic view, the document view (Figure 3(c)) is synchronized with the topic view to display relevant document information based on a user's interaction. For example, when a user selects a topic keyword in the topic view, the document view displays a set of document snippets that contain the selected topic keyword (Figure 3(c)). A user can also interact with the content displayed in the document view to constrain the topics shown in the topic view. For example, in an email analysis application, assume that a user is interested in only the emails exchanged between two specific people. S/he can constrain the sender and receiver parameters in the document view to focus on only the subset of emails that s/he is interested in. Accordingly, the topic view is updated to show only the topics relevant to the selected subset of emails. This function is quite useful, since it allows the users to flexibly analyze documents both top-down (i.e., from topics to individual documents) and bottom-up (i.e., from individual documents to topics).

The data navigation (Figure 3(a)) and action menu (Figure 3(d)) provide users with additional interaction tools to examine the document collection. For example, a user can interact with the topic legend to quickly view the derived topics in a linear list (Figure 3(d)) and then select to view the topics that s/he is interested in. Users can also interactively change the stacking order of the topic layers, or request the display of the topic strength in a numeric format (Figure 4). As explained later, these interactions not only allow users to better comprehend a text summary based on their needs, but they also help address certain technical challenges in both text analytics and visualization.

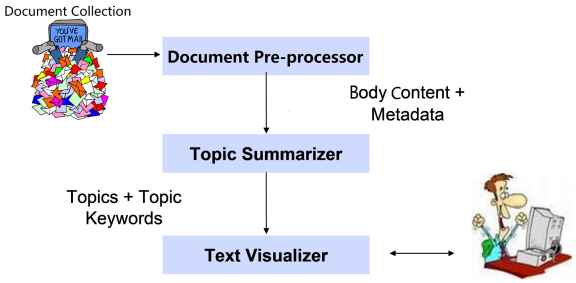


Fig. 5. TIARA architecture.

### 3.2. Architecture

Figure 5 provides an overview of the TIARA architecture. It has three main modules: document preprocessor, topic summarizer, and text visualizer.

The input to TIARA is a collection of text documents. The document preprocessor is tasked to extract the main document body text and various metadata. Take email analysis as an example, the metadata includes the sender, receiver, and timestamp information in each email. The output of the email preprocessor is a collection of "clean" document content with associated metadata. This output is then sent to the topic summarizer which automatically extracts a set of topics along with various probabilistic measurements, such as the probability of a topic existence and the probability of a document belonging to a specific topic. Given the results produced by the topic summarizer, the text visualizer transforms the abstract analysis results into a comprehensible visualization. Users can then interact with the generated visualization to further their analysis.

## 4. TOPIC-BASED TEXT SUMMARY

To produce an effective text summarization, we have examined state-of-the-art text summarization techniques. Latent semantic models appear very effective for topic modeling and analysis [Dredze et al. 2008; McCallum et al. 2007]. Among these models, previous experiments show that the Latent Dirichlet Allocation (LDA) model outperforms others significantly in text summarization [Dredze et al. 2008]. Therefore, we employ LDA in TIARA to summarize a text collection at two levels. First, we use it to automatically extract a set of latent topics for the whole text collection. Second, we use LDA to summarize each document in a set of keywords [Dredze et al. 2008]. To make the LDA-derived results more comprehensible to ordinary users, we also enhance the results from two aspects. First, we use a set of criteria to order the LDA-derived topics so TIARA can present the top-$N$ most meaningful topics to users first. Second, in each topic we identify the top-$K$ most salient keywords to describe the topic at each given time point.

### 4.1. Bilevel Text Summarization

Given a text collection, the LDA method automatically learns a set of topics. Formally, each topic is defined as follows.

*Topics and Topic Keywords.* A topic represents the thematic content common to a set of text documents. Each topic is characterized by a distribution over a set of keywords. We use the term *topic keywords* to refer to these keywords. Each keyword has a probability measuring the likelihood of this keyword appearing in the related topic.

*Topic Strength over Time.* Since TIARA currently focuses on helping users analyze how topics evolve over time, we measure the strength of a topic at a specific time point. Assume that a topic $z_k$ summarizes the thematic content over a set of documents $D$. Given time $t$, we can then compute the distribution of topic $z_k$ over the subset of documents with timestamp $t$.

$$S_t[k] = \sum_{d_m \in D(t)} L(d_m) \times p(z_k|d_m) \quad (1)$$

Here $d_m$ is the $m$th document in collection $D(t)$, which is the set of documents at time $t$. Function $L(d_m)$ computes the normalized length of document $d_m$, while $p(z_k|d_m)$ calculates the distribution of topic $z_k$ in $d_m$. In Gibbs sampling of LDA, we iteratively assign a topic label to each word. The topic distribution $p(z_k|d_m)$ for topic $z_k$ in document $m$ is calculated based on the count of words which are assigned with topic $z_k$ in document $m$, given the prior Dirichlet distribution [Blei et al. 2003].

Naturally, Eq. (1) defines the "strength" of a topic at time $t$. Intuitively, stronger topics are covered by more documents in a collection. Visually, the stronger topics appear taller.

*4.1.1. Summarizing a Document.* In addition to summarizing a collection of documents, we extract a set of *document keywords* to describe the gist of each document [Dredze et al. 2008]. Like a topic keyword, a document keyword is also associated with a score measuring the relevance of the keyword to the document [Dredze et al. 2008]. To summarize a document based on keywords, the most important keywords are selected by combining the document-topic distribution and topic-word distribution. Consequently, the words corresponding to the most important topic(s) in a document are more likely to be selected.

### 4.2. Enhancing Summarization Results

Although LDA is effective in deriving latent topics [Dredze et al. 2008], the raw results often cannot be directly visualized for two reasons.

First, the LDA output is statistically inferred (e.g., using Gibbs sampling) and the inferred results may not be useful for text analysis. Consider an attorney's email collection, where each email contains a disclaimer. Consequently, a learned topic is about the disclaimer[1]. Directly visualizing such a general topic does not help a user in his/her email analysis. It may even distract the user from seeing more useful information.

Second, LDA is a general model for learning topics and keywords without considering the unique characteristics of a text collection in a specific domain (e.g., an email collection). Thus, some of the learned topics may not be useful for text analysis. Consider an engineer's email collection that contains a large number of received ACM newsletters. It is highly likely that a derived topic is about the newsletters. However, one key goal for email analysis is often to understand *email conversations* among people instead of one-way messages like the ACM newsletters [Viegas et al. 2006].

To provide users with more meaningful summarization results, we have enhanced the LDA-derived results in two areas. First, we rank the LDA-learned topics by a set of criteria and then select the top-$N$ most meaningful ones to visualize first[2]. Second, we also extract the top-$K$ most salient keywords specific to a time point for each topic to depict the topic content evolution over time.

[1]One way to avoid deriving such a topic is to remove the disclaimer from every email. However, text processing is not always 100% accurate.

[2]TIARA automatically includes a topic in its visual summary if its importance exceeds a threshold. Otherwise, a user must explicitly request it.

*4.2.1. Topic Ranking and Selection.* Given a set of LDA-derived topics, our goal is to present users with the most meaningful topics first. However, the definition of meaningfulness varies from one user to another depending on their analysis tasks. For example, one user may want to see the most popular topics that cover most of the content in a text collection, while the other may prefer to see the most distinctive topics that cover very different content than others do. We thus have experimented with various topic ranking methods, each of which focuses on computing ranks based on one or more criteria.

As described next, we have used both application-independent as well as application-specific ranking criteria.

*Topic Ranking by Topic Coverage and Variance.* Our first method uses two application-independent criteria and is quite intuitive: finding "popular" topics that cover a significant portion of the corpus content. However, we prefer content variance in "popular" topics, since we are not interested in the topics that constantly appear in all the documents (e.g., a topic on disclaimer derived from a legal document collection). As a result, we use a weighted combination of content coverage and topic variance scores to measure a topic rank. Mathematically, we define the topic coverage and topic variance in Eqs. (2) and (3), respectively.

$$\mu(z_k) = \sum_{m=1}^{M} N_m \cdot p(z_k|d_m) \Big/ \sum_{m=1}^{M} N_m \tag{2}$$

$$\sigma(z_k) = \sqrt{\sum_{m=1}^{M} N_m \cdot \big(p(z_k|d_m) - \mu(z_k)\big)^2 \Big/ \sum_{m=1}^{M} N_m} \tag{3}$$

Here $z_k$ is the $k$th topic, $M$ is the number of documents, $N_m$ is the document length, and $p(z_k|d_m)$ is the document-topic distribution. $\mu(z_k)$ is a weighted average of a specific topic $z_k$ in different documents. The larger the score is, the better coverage of documents the topic is. Similarly, $\sigma(z_k)$ measures a weighted variance of topic distribution among different documents. A larger score indicates that the topic distribution tends to be more different for different documents.

Then the rank of topic $z_k$ is formulated as

$$R_k = (\mu(z_k))^{\lambda_1} \cdot (\sigma(z_k))^{\lambda_2}, \tag{4}$$

where $\lambda_1$ and $\lambda_2$ are control parameters, $\lambda_1 + \lambda_2 = 1$.

*Topic Ranking by Topic Distinctiveness.* While our first topic ranking method uses the representative power of a topic, our second method leverages the discriminating power of a topic. This method is mainly motivated by our own text analysis experience where users may be interested in topics that are distinctly different from one another so they can get a full understanding of a document collection. For this purpose, we develop a Laplacian score [He et al. 2005]-based topic ranking method that assigns higher ranks to topics with higher discriminating power. We base our method on the empirical observation that two similar documents most likely belong to the same topic, while dissimilar documents probably belong to different topics. Furthermore, the Laplacian score of a topic reflects its power in discriminating documents from different classes and preserving the local structure of a document collection. Our method consists of five main steps.

(1) Represent each document $d_m$ as a node in a graph. Its features are represented by a vector $\mathbf{v}_m$.

(2) Construct the $T$-nearest neighbor graph based on a similarity matrix $\mathbf{S}$ where $\mathbf{S}_{ij} = \exp\{-e_{ij}^2/c^2\}$. Here, $c$ is a constant and $e_{ij}$ can be either Euclidian distance or Hellinger distance [34] of document vectors $\mathbf{v}_i$ and $\mathbf{v}_j$.
(3) Compute graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{D}_{ii} = \sum_j \mathbf{S}_{ij}$ is the degree of the $i$th node.
(4) For each topic, $\mathbf{u}_k = (p(z_k|d_1), p(z_k|d_2), ..., p(z_k|d_M))^T \in R^M$ , let

$$\tilde{\mathbf{u}}_k = \mathbf{u}_k - \frac{\mathbf{u}_k^T \mathbf{D}\mathbf{1}}{\mathbf{1}_k^T \mathbf{D}\mathbf{1}}\mathbf{1}. \tag{5}$$

Here, $\mathbf{1} = (1, 1, ....1)^T$.
(5) Compute the Laplacian score of the $k$th topic.

$$L_k = \frac{\tilde{\mathbf{u}}_k^T \mathbf{L}\tilde{\mathbf{u}}_k}{\tilde{\mathbf{u}}_k^T \mathbf{D}\tilde{\mathbf{u}}_k} \tag{6}$$

*Topic Ranking by Topic Information Gain.* The amount of mutual information between two topics measures the amount of information they share. In other words, the amount of mutual information helps measuring how much knowing one of the topics would reduce the uncertainty of knowing the other. Using this metric, we can rank topics by the greatest amount of mutual information between any two topics. Specifically, we use the following procedure to determine the rank of each topic [Sahami 1998].

(1) For $\forall i, j$, first compute $MI(\mathbf{u}_i, \mathbf{u}_j)$ based on the document-topic distributions of $\mathbf{u}_i$ and $\mathbf{u}_j$. Then construct a complete graph $G$ where the weight of an edge $e_{\mathbf{u}_i,\mathbf{u}_j}$ is $MI(\mathbf{u}_i, \mathbf{u}_j)$.
(2) Build the maximal spanning tree MST of the complete graph $G$.
(3) Define the relevant topic set $R_t = \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_K\}$ and the corresponding edges in MST.
(4) While $|R_t| > 0$,
  (4.1) if $\exists k \in G$ is not connected to the others in $R_t$, remove this topic $\mathbf{u}_v$ ($R_t \leftarrow R_t - \mathbf{u}_v$).
  (4.2) otherwise remove the least weighted edge in $R_t$.
(5) Rank the topics according to the order in which they were removed. Rank the topic removed last the highest.

*Topic Ranking by Topic Dissimilarity and Redundancy.* Our last ranking method aims to maximize topic diversity and minimize redundancy. Unlike our preceding methods, which use topic-document relationships, this method employs topic-word distributions to compute topic similarity. We have modified the algorithm proposed in Mitra et al. [2002] to derive a topic rank.

(1) For $\forall i, j$, compute the similarity $s_{ij}$ for $\boldsymbol{\varphi}_i$ and $\boldsymbol{\varphi}_j$ based on the *maximal information compression index* [Mitra et al. 2002]. Here, $\boldsymbol{\varphi}_i = (p(w_i|z_1), p(w_i|z_2), ..., p(w_i|z_V))^T$, where $p(w_i|z_k)$ is the topic-word distribution, $V$ is the number of words.
(2) Sort the similarities for each topic.
(3) Define the reduced topic set $R_t = \{\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, ..., \boldsymbol{\varphi}_K\}$.
(4) While $|R_t| > 0$, remove $\boldsymbol{\varphi}_k$ in $R_t$ which satisfies $j = \arg\max_i \max_j s_{ij}$.
(5) The rank of a topic is determined by the topic removal order. The topic removed last should be ranked the highest.

*Topic Ranking by Application-Specific Features.* As described before, we have used a set of methods to rank topics by a set of application-independent criteria. Application-specific information such as email metadata, however, can also be very helpful in determining the meaningfulness of a topic. For example, if a topic mainly includes email messages that have never been read or replied to, this topic may be less important. In contrast, if a topic contains emails that are not only read but also frequently replied to, this could be an indication of an important topic that is worth further investigation. We thus also allow the incorporation of additional domain information in our ranking model. For example, in an email application, we define

$$r_k^{(reply)} = \sum_m^M p(z_k|d_m)(5 \cdot (\#\text{self reply})_{d_m} + 2 \cdot (\#\text{other reply})_{d_m}), \tag{7}$$

where $(\#\text{self reply})_{d_m}$ is the reply count by the email owner for document $d_m$, and $(\#\text{other reply})_{d_m}$ is the reply count by other people for document $d_m$. Based on this measure, the topics with more email replies will be ranked higher. To incorporate $r_k^{(reply)}$, we multiply this value with the ranking scores computed earlier.

*4.2.2. Topic Keyword Ranking and Selection.* Like the LDA-derived topics, the extracted topic keywords may not be very useful for users in their analysis tasks. For example, most of the topics extracted to summarize Enron CEO Ken Lay's emails contain the keyword "*Enron*". In addition, one of our goals is to help users analyze thematic content changes in each topic over time. However, the LDA-derived topic keywords are time insensitive. For example, a topic derived to summarize a conference activity may contain general keywords like "*meeting*" or "*conference*" over the *entire* life span of the topic. Repeatedly showing the same keywords across many topics or across the entire timeline of a topic would not help users identify unique topics or topic changes over time. Therefore, we rank topic keywords based on their importance to a topic and to a specific time frame.

Inspired by the term reweighting techniques used in Information Retrieval (IR) [Salton and Buckley 1988; Lan et al. 2005], we have experimented with two LDA-versions of the TFIDF-like scores to compute the importance of keyword $w_i$ in $z_k$ at time $t$.

$$KR_1 = \frac{\varphi_{k,i}}{\sum_{l=1}^K \varphi_{l,i}} \tag{8}$$

$$KR_2 = \varphi_{k,i} \cdot \log \frac{\varphi_{k,i}}{(\prod_{l=1}^K \varphi_{l,i})^{\frac{1}{K}}} \tag{9}$$

Here the native word weight $\varphi_{k,i} = p(w_i|z_k)$ generated by LDA corresponds to the term frequency. The topic proportion sum and topic proportion product in $KR_1$ and $KR_2$ are used respectively to simulate the inverse document frequency to reweight the native word weights. Here, $K$ is the number of topics.

*4.2.3. Experiments.* To compare the performance of all our topic and keyword ranking methods, we have conducted a series of experiments on two datasets: a personal email collection and a collection of news articles relevant to "*AIG insurance*". The first dataset is a personal email collection dated from February to December 2008 with 8326 email messages. Each email is associated with a set of metadata such as sender, receiver, time, subject, body, and reply counts. Only the subject and the body of each email were used to train the topic model. We preprocessed each email to remove irrelevant information such as email signature and also did stop-word removal. After

Table I. Email Topic Ranking Results without Reply History

| Retrieved | Top 5 | Top 10 |
|---|---|---|
| M.S. | $0.800 \pm 0.000$ | $0.620 \pm 0.028$ |
| L.S | $\mathbf{1.000 \pm 0.000}$ | $\mathbf{0.780 \pm 0.028}$ |
| M.I | $0.760 \pm 0.106$ | $0.740 \pm 0.035$ |
| W.S | $0.440 \pm 0.057$ | $0.480 \pm 0.028$ |

Table II. Email Topic Ranking Results with Reply History

| Retrieved | Top 5 | Top 10 |
|---|---|---|
| M.S. | $0.760 \pm 0.057$ | $0.640 \pm 0.035$ |
| L.S | $0.920 \pm 0.069$ | $\mathbf{0.900 \pm 0.000}$ |
| M.I | $\mathbf{0.960 \pm 0.057}$ | $0.860 \pm 0.035$ |
| W.S | $0.520 \pm 0.056$ | $0.560 \pm 0.035$ |

Note: "C.V." represents *Weighted Topic Coverage and Variation* method, "L.S." represents *Laplacian Score* method, "M.I." represents *Pairwise Mutual Information* method, and "T.S." represents *Topic Similarity* method.

Table III. News (AIG) Topic Ranking Results

| Retrieved | Top 5 | Top 10 |
|---|---|---|
| M.S. | $0.640 \pm 0.057$ | $0.68 \pm 0.028$ |
| L.S | $\mathbf{0.760 \pm 0.057}$ | $\mathbf{0.76 \pm 0.035}$ |
| M.I | $\mathbf{0.760 \pm 0.057}$ | $0.74 \pm 0.035$ |
| W.S | $0.720 \pm 0.069$ | $0.70 \pm 0.045$ |

preprocessing, the email collection contained 958,069 word tokens in total. The second dataset is an online document collection that contains text retrieved by a search engine. These documents came from various news, blog, and forum Web sites. The search engine used "AIG insurance" as the query and retrieved 34,690 documents from January 2008 to April 2009. After preprocessing, the final AIG collection contained 11,491,246 word tokens in total.

For the email dataset, the person who owned the email collection helped us annotate the results. She was asked to label each topic as either "very important", "somewhat important", or "unimportant". In addition, for each topic, she was also asked to label each of the top 50 keywords as either "relevant" or "irrelevant". When asked about how she ranked these topics, the email owner summarized her criteria as: (1) A "very important" topic clearly describes a major project that the email owner was heavily involved in. (2) A "somewhat important" topic focuses on a specific event, such as writing a paper. (3) An "unimportant" topic either lacks a clear focus or is about very general work-related activities. For the AIG news dataset, we asked a person who was familiar with the recent AIG-related events to help us annotate the topics and keywords. This person determined the importance of a topic as follows: (1) A "very important" topic clearly describes an event directly related to AIG, for example, the AIG bonus controversy. (2) A "somewhat important" topic focuses on some background events such as the 2009 presidential election or the financial market crisis. (3) An "unimportant" topic is defined as either confusing or irrelevant, for example, a topic about various advertisements.

Given the annotated topics and keywords, we compared the automatic topic and keyword ranking results with the human-provided results using the $F_1$-measure, a criterion commonly used in IR. Following the IR tradition, in our analysis we categorized our topics annotated by our experts into both "relevant" and "irrelevant". The "relevant" topics are those that are either "very important" or "somewhat important" while "irrelevant" topics are those that are "unimportant". Similarly, based on our topic or keyword ranking methods, each topic or keyword can be categorized as either "retrieved" or "not retrieved" depending on the assigned ranks and the cut-off threshold used in each evaluation metric ("top 5" means only the top five keywords are retrieved).

Table IV. Email Word Ranking Results

| Retrieved | Top 10 | Top 20 |
|---|---|---|
| $KR_0$ | $0.535 \pm 0.055$ | $0.442 \pm 0.048$ |
| $KR_1$ | $\mathbf{0.701 \pm 0.067}$ | $\mathbf{0.600 \pm 0.043}$ |
| $KR_2$ | $0.616 \pm 0.078$ | $0.551 \pm 0.049$ |

Table V. News (AIG) Word Ranking Results

| Retrieved | Top 10 | Top 20 |
|---|---|---|
| $KR_0$ | $0.466 \pm 0.111$ | $0.445 \pm 0.083$ |
| $KR_1$ | $\mathbf{0.662 \pm 0.094}$ | $\mathbf{0.614 \pm 0.054}$ |
| $KR_2$ | $0.403 \pm 0.079$ | $0.343 \pm 0.048$ |

The topic ranking results for the email dataset (without reply history) are shown in Table I. From the results, the *Laplacian score* method outperformed the rest. In particular, all the top-5 retrieved topics were labeled by the email owner as most meaningful.

Moreover, when we add the reply history, even though the mutual information-based ranking method performed the best at the top-5 retrieval level, the Laplacian score-based method performed comparably (Table II). The latter method also performed the best at the top-10 retrieval level.

The topic ranking results for the AIG dataset are shown in Table III. Again, the Laplacian score method outperformed all other methods. Overall, the Laplacian score ranking method seems to capture the essence of a meaningful topic the best.

Furthermore, we performed a set of experiments on combining multiple ranking methods to see whether a hybrid method could further improve the performance. It turned out the hybrid methods that we tested did not outperform the individual ranking methods. This is a nontrivial problem, since not only do we need to figure out which ranking methods should be combined, but also how to combine them. Fully addressing this problem requires much future work and is beyond the scope of this article.

By Eqs. (8) or (9), we rank topic keywords by their importance. We then select the top-$K$ ranked keywords to display at each time $t$. To validate these two keyword ranking methods, we have also conducted a set of experiments on the same two datasets used by our topic ranking experiments.

The keyword ranking results are shown in Table IV and Table V. In these tables, $KR_0$ is the baseline that uses the LDA estimated parameters directly; $KR_1$ and $KR_2$ are defined in Eqs. (8) or (9). We can see that $KR_1$ performs better than $KR_0$ and $KR_2$. It shows that for our two datasets, topic proportion sum is better than topic proportion production in weighing the proportion.

## 5. TOPIC-BASED VISUAL TEXT SUMMARY

As described in Section 4, the LDA output is a set of topics and keywords that summarize the thematic content of a text collection. Here we describe in detail how TIARA visualizes such output.

### 5.1. Overall Visual Design

To encode the derived topics and their changes over time, we have employed a stacked graph. We use colored layers to depict individual topics, including their time evolution, thematic content (topic keywords), and thematic strength (Figure 1). As described shortly, we have extended an ordinary stacked graph in two areas to achieve our goal. First, we optimize the stacking order of the topic layers to produce a semantically faithful and aesthetically pleasing stacked graph. Second, we fill each topic layer with

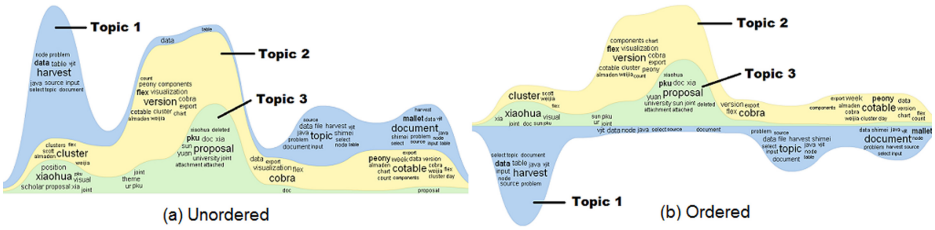


Fig. 6. Ordering of topic layers.

keyword clouds at different time points to informatively convey the content changes over time.

In general, there are four key steps in producing a stacked graph [Byron and Wattenberg 2008]: (1) computing the geometry of layers, (2) layer coloring, (3) layer ordering, and (4) layer labeling.

For the first step, we adopt the wiggle-minimizing method proposed in Byron and Wattenberg [2008] to create a smooth topic layer. In the second step, we use varied colors to differentiate adjacent layers. We also use colors to depict topic relationships. In particular, we consider the semantic properties of topics when making color choices. Currently, we use colors in the same family with varied hues to color layers that represent semantically similar topics. For example, Topics 1 and 2 in Figure 1 are both colored in blue with different hues, since they both talk about intellectual property. We measure the semantic similarity of two topics by counting the number of the same documents belonging to both of them.

$$SemSim(z_i, z_j) = C(D_i \cap D_j)/Max(C(D_i), C(D_j)) \tag{10}$$

Here $z_i$ and $z_j$ are the $i$th and $j$th topics, to which documents $D_i$ and $D_j$ belong, respectively. Function $C()$ counts the number of documents. The value is normalized to fall between [0, 1].

While the first two steps are straightforward, the last two require special efforts to handle the complexity of encoding our data.

### 5.2. Ordering Topic Layers

The stacking order of the topic layers directly impacts the legibility and aesthetics of a stacked graph [Byron and Wattenberg 2008]. In our case, the distortion caused by undesirable layer ordering can also diminish the usable space for displaying the topic keywords within a topic layer. Due to the distortion, the usable spaces within these three topic layers diminish. This latter limitation is critical to TIARA, since it needs to encode a rich amount of content (topic keywords) within each topic layer.

Our initial experiments also found that ordering topic layers purely based on aesthetic criteria proposed by Byron and Wattenberg [2008] is insufficient for text analysis. Topic ordering without considering the semantic properties of topics may reduce the gestalt effect of perceiving desired data associations, which in turn hinders users from discovering data patterns. In Figure 6(b), the bottom two topics are closely related, describing two complementary efforts. The complementary evolution of the two topics can be easily observed in Figure 6(b). For example, during the time interval t, Topic 3 becomes increasingly more active while Topic 1 is dormant. However, such a relationship would be hard to perceive in Figure 6(a) as the two topics are not placed next to each other.

To create an aesthetically appealing and semantically meaningful visual text summary, we extend the layer ordering method in Byron and Wattenberg [2008] to meet

three criteria: (1) minimizing the layer distortion, (2) maximizing the available space within each layer to accommodate rich thematic content, and (3) ensuring visual proximity of layers to be proportional to their semantic similarity (Eq. (10)). Simultaneously satisfying all three criteria is an optimization problem. Currently, we use a three-step greedy algorithm to approximate it.

First, we compute the volatility of a topic layer $z_i$ based on the standard deviation of the layer heights (Eq. (11)). This metric states that a topic layer is "volatile" if its strength fluctuates greatly over time.

$$V(z_i) = F_\sigma(h_{ik}) \tag{11}$$

Here, $h_{ik}$ is the layer height at time $t_k$; $F_\sigma$ computes the standard deviation of the layer heights.

Second, we sort all the topic layers by their volatility and start time. The least volatile one with the earliest start time is placed in the center of the graph. Third, we select the next topic layer to add in by an "inside-out" order [Byron and Wattenberg 2008]. The next topic is selected based on four properties: start time, volatility (Eq. (11)), semantic similarity with the previously added topic (Eq. (10)), and geometric complementariness with the previous topic (Eq. (12)). Step three is repeated until all the topic layers are added to the stacked graph.

$$GC(z_i, z_j) = F_\sigma(h_{ik} + h_{jk}) \tag{12}$$

Consequently, our approach balances all three layer-ordering criteria. First, it places the "flatter" topic layers toward the center of the graph and curvier ones along the edge to minimize the layer distortion [Byron and Wattenberg 2008]. Second, it neighbors geometrically complementary topic layers to maximize the usable space within each layer. Third, it groups semantically similar topic layers together to facilitate topic association and comparison. In addition, we tested these two methods in our further deployment; most test cases show that our method outperforms the method in Byron and Wattenberg [2008].

### 5.3. Placing Topic Keyword Clouds

Unlike typical stacked graphs, which mainly illustrate numeric trend changes with a few labels [Byron and Wattenberg 2008; Havre et al. 2002], TIARA is designed to convey thematic content changes. In other words, the label of each layer is no longer a short text string but multiple sets of keywords, abstracting the content of a topic at different time points. Displaying these keywords in a topic layer is nontrivial, since it must satisfy multiple, potentially competing constraints. For example, we want to display as many keywords as possible to informatively describe a topic while preventing keyword overflowing across the topic boundary.

Currently, our keyword placement method considers three factors: (1) temporal proximity, (2) content legibility, and (3) content amount. The first factor ensures that topic keywords be placed near the most relevant time coordinate. The second criterion requires that keywords be legible, such as avoiding keyword occlusions and overflowing across topic boundaries. The third criterion attempts to maximize the use of available space in a topic layer to display as many keywords as allowed. To meet all three criteria, we develop a two-step algorithm to place topic keywords as a series of keyword clouds along the timeline within a topic layer.

First, we sort all keyword sets of a topic by their timestamp. Starting from the earliest timestamp, we try to find a suitable space to place the associated keyword set within the topic layer. Second, we pack the keywords as a keyword cloud in the found space.

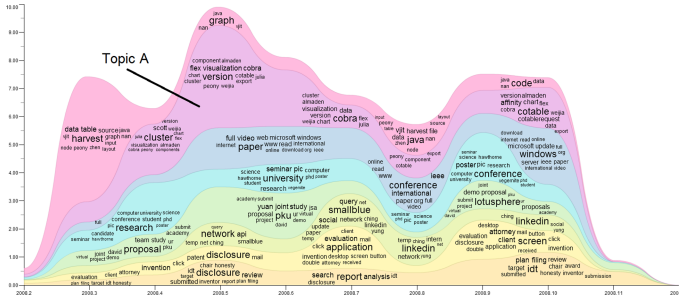


Fig. 7. Alternative layout of topic layers.

*Space Allocation.* To locate a suitable space for placing a set of topic keywords at time t, we search the neighborhood of $t(t \pm \sigma)$. Let $\delta$ be the time unit ($\delta = t_{i+1} - t_i$) and $\sigma < 0.5\delta$ to ensure that the keywords be placed near $t$. To ensure the legibility of topic keywords, TIARA requires that a keyword be displayed in a minimal legible font size or bigger. When evaluating the neighborhood of $t$, there are three possible outcomes.

First, if there is adequate space to fit a set of keywords ($K > 3$) in the neighborhood of time $t$, TIARA marks the space.

Second, if there is no adequate space within the allowed neighborhood to fit a single keyword in the minimal legible font, TIARA merges these keywords with those associated with time $t+\delta$. It then advances to the time point $t+\delta$ to look for suitable space. If TIARA is still unable to find a suitable space, it will drop the keywords at $t$ to ensure temporal proximity of the keyword placement (criterion 1). Otherwise, TIARA marks the space found near $t + \delta$.

Third, TIARA finds adequate space to fit only a few keywords ($K <= 3$). To avoid visual clutter, TIARA tries to minimize the number of scattered keyword clusters within a topic layer. It thus looks ahead to check the space around the time point $t + \delta$. If there is ample space (area > f), TIARA merges the keywords at $t$ and $t + \delta$ and places them near $t + \delta$. For example, TIARA merges Topic As keywords for March with those of April and places them near April (Figure 7). If both usable areas near $t$ and $t + \delta$ are small (i.e., each space fits fewer than 3 keywords), TIARA then combines the two areas to fit the merged keywords.

The preceding steps are repeated until every keyword set is processed.

*Keyword Cloud Packing.* Once TIARA finds a suitable space, it creates a keyword cloud in the allowed space. Our approach is based on Wordle[3] to pack keywords as tightly as possible. For computational efficiency, however, we consider only spaces surrounding a packed word but not inside it (i.e., the space available between or within characters). Like Wordle, TIARA places the most important keywords in the center. The keyword's font size is proportional to its importance but cannot be smaller than the minimum legible font size. To avoid keyword "overflow", we adopt a particle-based method [Luboschik et al. 2008] to pack the keywords within the boundary of a topic layer. We pack as many keywords as possible in a designated space as long as every displayed keyword is legible and none spills over the topic boundary.

[3]www.wordle.net

Besides distributing keyword clouds along time within a topic layer, we also provide a tool tip to show all keywords associated with the topic (Figure 1). This view offers users a topic overview regardless of its thematic changes.

### 5.4. Alternative Topic-Layer Layout

Initially, we laid out topic layers by an "inside-out" order [3]. While this layout produces a smooth and symmetric appearance (Figure 1), the symmetric layout leads to two artificial groups of topics, one on each side of the x-axis. This perception may be caused by the rich textual information displayed in each layer. To amend this situation, we then experimented with an alternative layout where all topic layers are stacked on the one side of the x-axis. In this layout, flatter topic layers are placed near the bottom and curvier ones on the top. The resulting visual summary eliminates the artificial grouping effect and also produces a more compact layout (Figure 7). To assess the two different designs, we have consulted four visual/graphics designers. While one preferred the original design (Figure 1), three favored the other (Figure 7). Their rationales are similar to the ones described earlier. As a result, we decided to use the design where all topic layers are stacked on one side of the x-axis in TIARA.

## 6. TOPIC-BASED VISUAL TEXT ANALYSIS

To better aid users in consuming complex text summarization results and perform deeper analysis, TIARA allows the users to interact with the generated visual summary and examine relevant data from multiple perspectives. Specifically, we support several types of visual interactions.

*Topic Details on Demand.* Due to limited screen real estate, often only a subset of topic keywords can be displayed within a topic layer. Moreover, the narrower a topic layer, the fewer topic keywords it can accommodate. In such cases, users may find the displayed information inadequate for their tasks and would want more details about the topic. To support such requests, TIARA allows users to interactively zoom in on a selected topic or topic segments. Currently, TIARA uses a fisheye view technique [Sarkar and Brown 1994] to display more details of a selected topic (Figure 10).

*Text Information on Demand.* Our keyword-based summarization sometimes may not provide sufficient information for text analysis. It is thus often necessary for a user to examine the meaning of topic keywords in the context of the original text documents. In TIARA, users can select any keyword displayed in a topic layer to retrieve the relevant documents. Instead of displaying the matched documents in full, TIARA first displays the matched snippets (Figure 9). Currently, the snippets are created based on two criteria: keyword match and diversity. The first criterion ensures that a snippet contains the user-selected keyword. Since a document may contain the selected keyword in multiple places [Clarke et al. 2008], our second criterion states that a different snippet be used to represent a document if a similar snippet is already in use for representing another document. Consequently, TIARA provides diverse information for users to assess the meanings of interested keywords. The user can also interact with a snippet to view the full document.

*Coordinated Multiview Analysis.* Although topic keywords represent the gist of a topic, they may be insufficient for users to fully grasp the meaning of the topic due to the terseness and potential ambiguities of the keywords. Intuitively, this problem may be alleviated by using n-gram instead of unigram topic keywords. However, our experiments show that the quality of n-gram-based summarization largely depends on the characteristics of a text corpus. For example, our n-gram-based email summarization

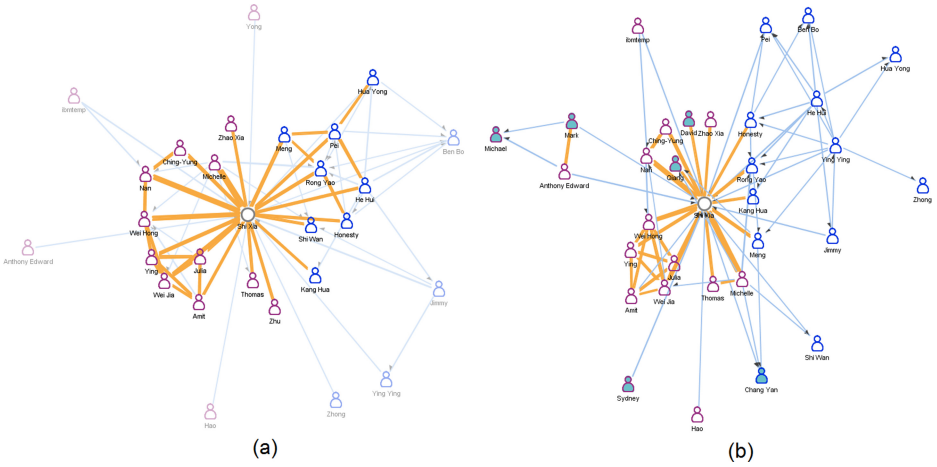


Fig. 8. Network diagrams showing sender and receiver relationships. (a) Highlighting the people involved in two-way communications. (b) Snapshot of the network at a different time point. Icons with blue fills are the "new" additions.

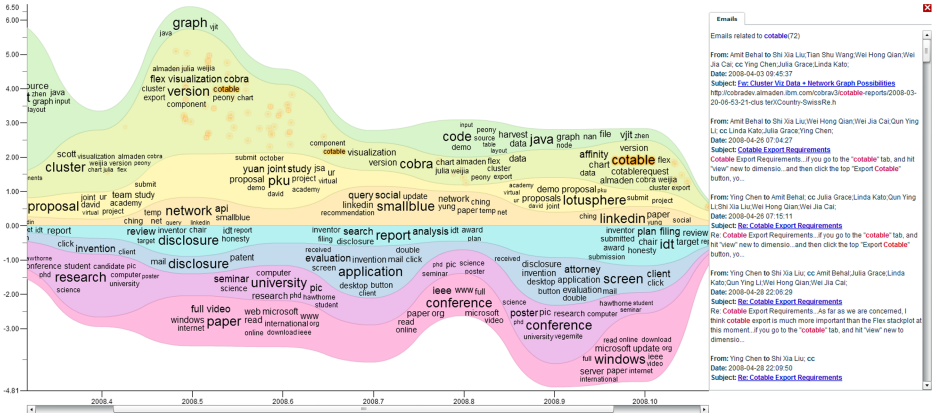

Fig. 9. Examining the email snippets that matched topic keyword "cotable".

appears ineffective since most n-gram keywords in emails represent people names or locations (e.g., New York).

As a corpus-neutral solution, TIARA allows users to interactively request additional data to help comprehend a topic. Currently, a user can request relevant metadata from a visual summary. In email analysis, users can request metadata such as sender and receiver information (e.g., Figure 8), while in company reputation analysis, they can ask for author and news source information. The visualizations of metadata are coordinated with the visual summary. As a result, users can use multiple views simultaneously to perform their analyses.

*Customizing Visual Analysis.* The major goal of TIARA is to encode rich amounts of information to aid users in their complex text analysis tasks. However, we must balance the amount of information that we could provide and the amount of information that a user can consume. If we provide too little information, the user may have little to go on. On the other hand, if we provide too much information, the user may be

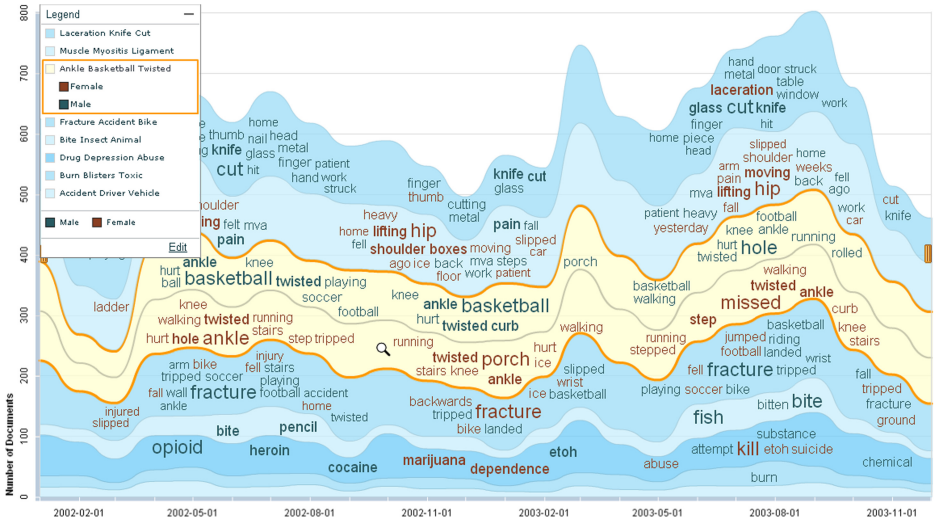


Fig. 10. TIARA's visual summary of the "cause of injury" field of the 23,000+ emergency room records from 2002 to 2003. In addition to visually summarizing the free-text fields (e.g. "cause of injury", "diagnosis", and "reason for visit" in the figure), TIARA incorporates structured fields (e.g. "patient gender") to explain the visual summary.

overwhelmed. To balance this, we provide several interaction tools that allow a user to manipulate the visualization and customize the amount of information s/he would like to view and analyze.

First, TIARA provides an interactive topic legend (Figure 3(d)) to allow a user to select the topics to be visualized. This tool is very useful especially if TIARA must deal with a number of topics for a huge collection of documents. In such cases, TIARA can only show a limited number of the topics at a time due to limited screen real estate. Using this tool, users can interactively choose which topics to display/hide based on their interests.

Second, TIARA provides users with an interactive topic layer ordering tool that allows users to interactively modify the stacking order of topics. Using this tool, users can interactively adjust the placements of the topic layers to suit their own information needs (e.g., facilitate topic comparison). Moreover, it helps compensate for the imperfection in our automatic layer ordering method (Section 5.2).

Third, TIARA provides tools for users to acquire additional information on demand. For example, when examining the topics in a visual summary, a user may want to compare the topic strengths at different time points. Instead of showing the numeric topic strengths at all time points in a summary, a user can interactively request the numeric topic strength at a selected time point (Figure 4). To facilitate the comparison of topic strengths, TIARA also allows the user to specify two time points to display the respective topic strengths simultaneously.

## 7. APPLICATIONS

We have applied TIARA to several real-world applications, ranging from email analysis to customer opinion analysis. Here we highlight our two key applications. Our first application is email analysis. In email analysis, TIARA is used to analyze both personal and public email corpora (e.g., our own inboxes and Enron email corpus). The goal for email analysis is to understand the thematic content of an email collection and how the thematic content evolves over time (Figure 1). Our second application is

patient record analysis. The goal of this application is to examine a number of emergency room patient records to answer a set of questions, including what the major causes of injury are and what the correlations are between the cause of injury and patient gender (Figure 10).

### 7.1. Visual Email Analysis

Given a set of emails, TIARA starts by creating a visual email summary (Figure 1). A visual email summary encodes multiple pieces of information, including the derived, top-N most important topics and the topic keywords associated with each topic (e.g., Figure 1). Given such a summary, a user can then examine a topic from several aspects. First, one can view topic changes in the form of keyword clouds distributed over time. In Figure 1, the top-most topic (green one) talks about "harvest, table, data..." in March, while discussing "java, code, vjit..." in August. Based on the varied heights of a topic at different time points, one can also observe how a topic's strength (activeness) changes over time. For example, the "harvest" topic is very active in March of 2008 but less active from April to July. To get the gist of a topic, one can request to view all the topic keywords in a tool tip (Figure 1).

From a visual email summary, a user can drill down to more detailed information. Assume that a user is interested in finding out more about the selected topic segment around April (Figure 2). She uses a magic lens [Bier et al. 1993] to obtain the senders/receivers of the emails belonging to the selected topic(s). Since the lens provides only the names of the senders and receivers, it lacks the detail as to how these people are related to each other (e.g., frequency and patterns of email exchanges). To obtain such information, the user can click on the lens to bring up a detailed network diagram (Figure 8). The owner of the email corpus is always placed in the center. There are two types of links: blue ones representing one-way communications and orange ones encoding two-way communications. The closer a person is placed to the owner, more frequent email exchanges occur between the two.

The user can interact with the network diagram and trigger the update in the visual text summary. For example, the user can select a subset of people (Figure 8(a)) and request a visual summary of the emails among only those people and the owner. Similarly, the user can view the relationship changes between the senders and receivers at different time points. Accordingly, the updated network view (Figure 8(b)) will also trigger the update of the visual text summary (e.g., summarizing only emails among the people shown in Figure 8(b)).

To interpret the meaning of a keyword in a topic, a user can retrieve emails that contain the keyword. Figure 9 shows the retrieved email snippets that match the user-selected keyword "cotable". By clicking on an email snippet, the user can then read the full email message. To provide users with the gist of an email, TIARA also summarizes each email in keywords. A user can also interactively constrain the emails s/he may be interested in (e.g., entering a keyword in the search bar to retrieve a set of matched emails).

### 7.2. Visual Patient Records Analysis

In our second application, we use the patient records available in NHAMCS (National Hospital Ambulatory Medical Care Survey) to analyze hospital emergency room situations. The current dataset includes 23,000 patient records from 2002 to 2003.

A patient record consists of multiple data fields, including free-text fields such as "diagnosis", "reason for visit", and "cause of injury", and structured value fields like patient gender and age. To facilitate analysis, a user can use the facet navigation panel to select and examine the data fields of his/her interest. Assume that the user

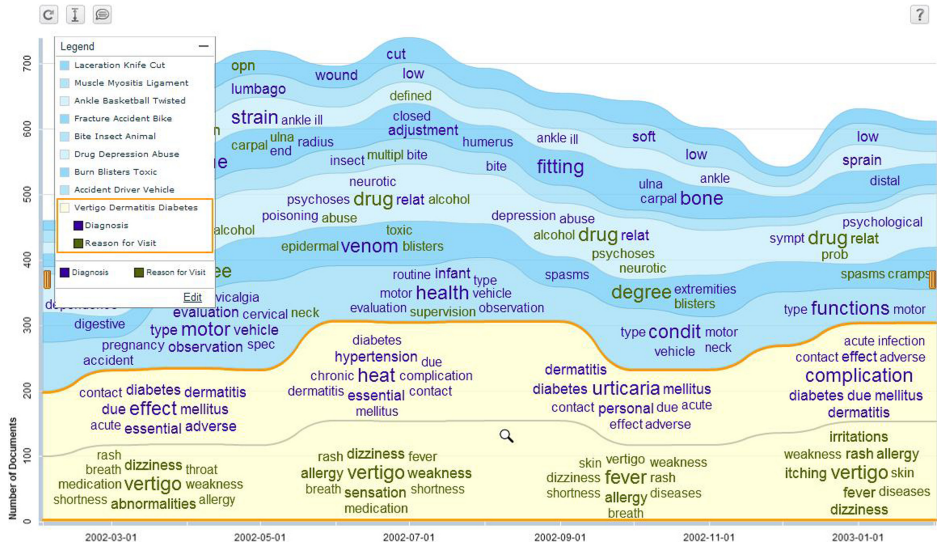


Fig. 11. Correlation analysis of two text fields "reason for visit" (in moss green) and "diagnosis" (in blue).

is interested in examining the "cause of injury" and its relation to "patient gender". TIARA first creates a visual summary of the patient records by their "cause of injury". Figure 10 shows the top 8 out of 15 derived topics.

The user can also zoom into a specific "cause of injury" (the one highlighted in Figure 10) to examine its relation to gender in detail. As shown in Figure 10, the selected topic layer is enlarged using a fisheye technique. The layer is also divided into two sublayers to display keywords by gender in two distinct word colors. As revealed by this display, the "cause of injury" for male patients tends to be sports-related, such as playing football or basketball. In contrast, female patients often injure themselves while performing routine activities, such as walking on stairs.

Similarly, TIARA can visually summarize and correlate any two free-text fields over time. For example, the user can select two text fields, "diagnosis" and "reason for visit" to examine their correlation along time. Figure 11 is an example of such a visual correlation summary. In this summary, the user focuses on the topic at the bottom (in yellow). A fisheye view is applied to enlarge the selected layer and split it into two sublayers, representing the two text fields. It is interesting to note that while the main "reason for visit" is "vertigo" throughout the year, the "diagnosis" differs. Specifically, in Spring, the diagnosis states that vertigo was caused by adverse effect of drugs, diabetes, or hypertension. However, in Summer, "vertigo" was attributed mainly to heat exhaustion.

### 7.3. Deployment

We have deployed TIARA in two ways: as a Web service or as a desktop application. To deploy TIARA as a Web application, we hosted an application server using Tomcat as the Web container for each application. Two Web services, one for email summarization and one for patient record analysis, have been set up and users can access TIARA via standard Web browsers like Internet Explorer. Currently, we only provide access to our company employees because we cannot host external servers. In this setting, users can only explore the datasets we provided.

To deploy TIARA as a desktop application, we implemented TIARA as a Plugin for IBM Lotus Notes, an enterprise email solution developed by Lotus. The TIARA Notes Plugin was deployed on an IBM internal software hosting service site and it was made available to all IBM employees worldwide. After downloading and installing the TIARA Notes Plugin, users can directly launch TIARA in Notes to analyze their emails.

So far, over 1000 downloads have been recorded. We also received many comments from users, such as "*I was very impressed by the way it deals with unstructured data", "I like the evolution graph with tag clouds on it*", "*The most impressive feature TIARA is its dynamic query and graphics rendering capability.*", "*TIARA visualization provides a quick overview of the documents being examined, which enables me to quickly find the active topics. Its cool!*".

## 8. EVALUATION

To evaluate the usefulness of TIARA, we have designed and conducted a series of formal studies by using within-subjects designs. Our studies aimed to evaluate how TIARA helps users perform realistic analysis tasks. We were also careful in our design so that users could accomplish such a task within a reasonable time span (i.e., within 30 minutes).

We have applied TIARA to two types of email analysis. One is to let our IBM colleagues use TIARA internally to analyze their own email archives. The other is to apply TIARA to analyze the Enron email corpus[4], examining the emails of Enron key personnel. Due to our limited knowledge of the Enron businesses, we found it difficult to design realistic email analysis tasks that can be achieved using the Enron corpus within a short time span (i.e., 15–30 minutes). Thus, we decided to evaluate TIARA using the email data contributed by our IBM colleagues.

### 8.1. Study Setup

To evaluate the effectiveness of TIARA in support of email analysis, we designed and conducted a comparison study. Our study compared the usefulness and usability of TIARA for email analysis with that of a baseline system. We chose Themail [Viegas et al. 2006] as our baseline mainly for three reasons. First, both TIARA and Themail focus on exploring the combined power of text analytics and interactive visualization. Second, both are tailored for email analysis. Third, we were able to obtain Themail code to run our experiments.

*8.1.1. Task Design.* We worked with the email owner to identify a set of email analysis tasks which she often performed in her work. Together we designed three types of email analysis tasks. The first type of task was the easiest. It required users to answer a set of specific questions using the email correspondences between just two people. We designed this set of tasks specifically to evaluate the effectiveness of TIARA in support of simple analysis tasks like those supported by Themail [Viegas et al. 2006]. The second type of task is more difficult, asking users to answer a set of questions about a specific event, which often involves email exchanges among multiple people. The users were provided with several clues that could be used to start the investigation. Finally, the third set of tasks was the most difficult as it provided few clues to start with. This type of task was intended to evaluate TIARA's effectiveness in more complex analysis tasks.

We designed a total of 24 questions for three types of tasks. Table VI shows a set of sample questions in each type of task. To ensure the validity of the tasks in the context

[4]http://www.cs.cmu.edu/ enron/

Table VI. Sample Questions in Our Study Tasks

| |
|---|
| **Task1 examining emails between two people** |
| *What are the three main concepts mentioned during their June emails?* |
| *Which month of 2008 is the most active in their email exchanges?* |
| **Task2 examining emails about a project named Cobra** |
| *Who were involved in Cobra?* |
| *What was discussed during the active period?* |
| **Task3 examining emails in general** |
| *What was the most active topic in May?* |
| *Who were the people involved in this topic?* |

of TIARA, we asked the email owner to answer all the questions using TIARA to obtain ground-truth answers. We amended problematic questions based on the email owner's experience. To accurately measure user performance (e.g., answer time) and avoid potential biases (e.g., users' communication capability), we used the obtained answers to phrase each question as a multiple-choice question. For example, when asked "what was the most active topic in May", a user was presented with four possible answers:

(a) visualization, analysis, social; (b) disclosure, iplaw, evaluation
(c) proposal, pku, yuan; (d) tan, offering, recruiting

*8.1.2. Method.* From the email corpora contributed by our IBM colleagues, we selected one colleague's email archive, consisting of about 10,000 emails in the year of 2008. We recruited ten users who had never used Themail and TIARA before. While five of them were familiar with the email owner or her work, another five were not. All these users majored in computer science and had used an email system (e.g., Lotus Notes, Microsoft Outlook) daily for at least 5 years. 70% were male and 30% were female. Each user was asked to perform all the tasks. To avoid potential biases such as learning effects, we designed two sets of similar but not identical questions for each task. For example, for task 1, examining emails between two people, we asked the user to examine the email communication between (1) Weijia Cai and the email owner; (2) Nan Cao and the email owner. Moreover, we randomly permuted the order of tasks and system usage.

At the beginning of each user session, we gave a brief tutorial of both systems. Each user was allotted 15 minutes for each task. The user was asked to fill in an evaluation survey after each task. We logged the users' answers to all questions and recorded all user interactions and the time used for answering each question.

To compare the performance of the two systems, we used both objective and subjective measures. Our three objective measures were derived from the log data: answer completion rate (percentage of the answered questions), answer error rate (percentage of incorrect answers), and answer time (the time spent on each question). For each system usage, we computed the mean answer completion rate, error rate, and answer time across all users and questions. Our three subjective measures were extracted from the user surveys: usefulness (how useful a system is for the task at hand), usability (how easy to use a system for the task at hand), and system satisfaction. All the subjective measures were rated on a 5-point scale, with 1 being the worst and 5 being the best. Furthermore, we asked the participants to indicate their most and least liked features of TIARA, and suggest additional features to enhance email analysis.

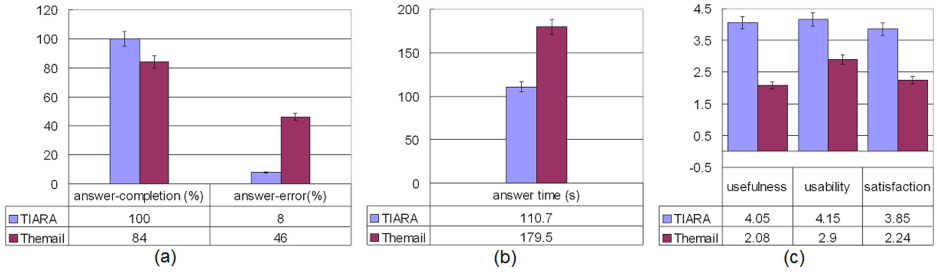


Fig. 12. Objective (a); (b) and subjective; (c) study results.

### 8.2. Results and Analysis

We examined both objective and subjective data collected from our study. As shown in Figure 12, TIARA outperformed the baseline system across all metrics. Objectively, TIARA helped users answer more questions, answer them more accurately, and in less time (Figure 12(a) and (b)). Subjectively, all users favored TIARA in terms of its usefulness and usability for the tasks they performed (Figure 12(c)).

*8.2.1. Objective Measures.* Compared to the baseline (Themail), TIARA performed significantly better on all three objective measures. First, the mean answer completion rate across all tasks and users was increased from 84% (Themail) to 100% (TIARA), a 19% increase (Figure 12(a)). Second, the mean answer error rate was reduced from 46% (Themail) to 8% (TIARA), an 83% reduction (Figure 12(a)). Third, the mean answer time over all questions and users was reduced from 179.5 secs (Themail) to 110.7 secs (TIARA), a 38% enhancement (Figure 12(b)). T-tests showed that all three improvements were statistically significant ($p < 0.001$ in all three measures).

We analyzed how various factors, such as task type and system usage, might have impacted the three measures. ANOVA tests found that both task type ($p < 0.001$) and system usage ($p < 0.001$) had influenced the answer completion and error rate significantly. A post hoc test further identified that TIARA had significantly improved the answer completion rate and error rate for more complex tasks (tasks 2 and 3 in Section 8.1.1), verifying our hypothesis. In addition, ANOVA tests found that three factors, task type ($p < 0.001$), system usage ($p < 0.001$), and users' familiarity with the email owner's work ($p < 0.001$), had impacted the answer time significantly.

To better understand these results, we further examined the nature of the tasks and the system usage. From our observations, more difficult tasks like identifying the most active topics for a particular time frame required users to have a good understanding of the entire email collection. Since Themail focused on depicting email correspondences between two people at a time, it was difficult for the users to obtain an overview of all the emails. In contrast, TIARA's visual summary allowed the users to gain a quick overview of the email collection. We also observed that participants who were familiar with the email owner's work often used their knowledge to eliminate impossible answers. This helped explain why these users answered questions faster but not necessarily more accurately.

For Task 1, it was interesting to observe that TIARA outperformed Themail in both answer completion rate (100% versus 95%) and error rate (0% versus 20%). However, Themail outperformed TIARA in answer time (171.55 secs versus 194.65 secs). After examining the recorded user interactions, we were able to explain the results. This task required users to analyze the emails between two people. Themail provided users with just the right amount of information for the task. In contrast, TIARA provided a summary of all emails, costing users more time to filter out irrelevant emails. However,

the simple TFIDF-based keyword display in Themail was not powerful enough for users to easily identify topics and observe thematic changes even within the emails between two people. Thus, TIARA was able to help users answer more questions and more accurately.

To better understand these results, we further examined the nature of the tasks and TIARA's usage. From our logs, we first investigated why users spent the most time on the simplest task (Task 1). Task 1 required the users to study email correspondences between two people. However, TIARA provided the summary of the whole email set, which required the users to locate the relevant topics first before obtaining answers. The more difficult tasks like Tasks 2 and 3 required the users to have a good understanding of the entire email collection. TIARA's visual summary precisely provided the users with a quick understanding of all emails. Moreover, we observed that the users used the provided clues to jump start their investigations. For example, Task 2 asked questions about a project named "Cobra". Almost all users started by finding the topics relevant to "Cobra" (containing topic keyword "Cobra") in the presented visual summary. After locating the relevant topics, they then examined the emails under these topics. In contrast, Task 3 provided little clues for the users to start with. They had to examine each topic in the summary before coming up with the answers. This perhaps explained why the users finished Task 2 the fastest.

Although our users took the longest time in finishing Task 1, they had completed the task perfectly. In contrast, they made small mistakes in both Tasks 2 and 3. When examining these mistakes, we found that most of the mistakes were related to questions requiring careful detailed analysis. For example, one such question was asking the users to identify a list of people involved in a project. Although candidate lists were provided as multiple choices, the users still needed to examine the details of relevant emails before making a selection. Currently, TIARA does not automatically extract the relationships between people and events. It thus did not directly help our users in such cases.

*8.2.2. Subjective Measures.* TIARA also outperformed Themail in all three subjective measures (Figure 12(c)). T-tests showed that the differences are statistically significant: usefulness ($p < 0.001$), usability ($p < 0.001$), and satisfaction ($p < 0.001$). Moreover, the task type ($p < 0.05$), system usage ($p < 0.001$), and users' familiarity with the email owner's work ($p < 0.001$) were the main factors that significantly impacted the usefulness. In addition, system usage ($p < 0.001$) was the main factor that had significantly impacted system usability and satisfaction.

When asked about their opinions of TIARA's key features, nine out of ten participants (90%) indicated that they liked the visual email summary the best. Seven out of ten (70%) participants stated that they least liked TIARA's inability to connect the email topics with the people involved in these topics.

To better understand the rationale behind their preferences, we further studied the participants' comments. Their comments were consistent with our analysis. Overall, TIARA was favored for two main reasons. First, TIARA's visual summary helped users gain a quick understanding of the underlying emails. Almost all users (90%) commented on how easily they could use TIARA to identify topics and observe content changes over time. In contrast, they expressed how unfit Themail was for answering questions in Tasks 2 and 3 (Section 8.1.1). Furthermore, users who could use their knowledge to quickly find the desired information in a visual summary tended to value TIARA more. For example, one user commented: "There is a lot of information (in this summary). But I know these (pointing to some topics) are irrelevant, so I focus on...". This helps explain why a user's knowledge had an impact on the usefulness measure. Second, users liked TIARA's interaction tools that provided them the flexibility to

examine email data from multiple angles. For example, one user commented, "... I did not know what these keywords meant. But it is good that I could look them up in the emails..." We also observed that users switched often between a visual summary, the email snippets, and the full email messages to glean information.

*8.2.3. Discussions.* During our study, we asked the participant's preferences on the two layouts, symmetric (Figure 1) and asymmetric layout (Figure 7). Five out of ten (50%) users preferred the asymmetric layout. They considered the layout more "natural, and liked that it packed the topic currents tightly for easy comparison. In contrast, they perceived a "divided" visualization in a symmetric layout, consisting of two parts, one on each side of the X axis. They considered the "division" an extra burden for comprehending the graph. Three (30%) users favored the symmetric layout. They felt the visualization was less distorted and thus more topic content could be shown. The remaining two users (20%) had no preference on the layout.

The participants' feedback also revealed the current limitations of TIARA. Eight of ten (80%) users expressed the need to provide feedback through visual interaction to improve the text analytics. For example, Topics 1 and 2 in Figure 1 should really be one topic. Several users wished that they could merge them together to avoid confusion. While visually merging the two topics may be easy, it is difficult to feed the merged results back to the LDA engine to enhance its future performance. Because of the terseness of topic keywords, several users also suggested that TIARA label each keyword with its semantic category to visually distinguish various concepts (e.g., people versus place). Since this feature involves named entity recognition, a research topic itself, we are exploring how to combine it with the LDA method used in TIARA.

Several users also expressed their desire to examine additional topics instead of just the top-N most important ones. Scaling a visual summary to accommodate a number of topics is nontrivial for several reasons. First, we need to properly label the topics so they can be easily selected. However, it is difficult to uniquely label an LDA-derived topic. We are exploring how to use a small set of keywords to label such a topic. Moreover, topic currents are ordered by three criteria (Section 5.2). Since the screen real estate is always limited, at a certain point we must minimize or hide displayed topics in order to show new topics. In such cases, it is unclear whether we should reorder the entire stack of topics or just incrementally order the new ones for the sake of performance and visual continuity.

## 9. CONCLUSIONS

In this article, we present an interactive, visual text analysis system called TIARA. TIARA supports both top-down and bottom-up visual text analysis. In a top-down process, TIARA provides a user with a visual summary of a text corpus. The user can drill down to specific text snippets and original text messages to better understand the summarized topics and topic keywords in context. In a bottom-up process, a user starts with a small set of text documents. Based on their content, the user may be interested in analyzing an expanded set of text documents. Accordingly, TIARA creates a visual summary of the expanded set. From the created visual summary, the user can then start another round of analysis.

As a result, TIARA provides users with three distinct benefits. First, it offers several criteria to enhance the topic modeling results to make them consumable by users. Second, its visual summary provides users with a quick overview of a large collection of text information, which helps them understand the thematic changes over time, and enables them to make rapid decisions on where they need to dig deeper. Third, TIARA offers users with a flexible set of interaction tools that help them digest text summaries in context, and examine relevant text information from multiple angles to

compensate for the deficiencies of current text summarization technology. Our application examples and preliminary evaluation also demonstrate that TIARA effectively aids users in their text analysis tasks, especially in the more complex tasks.

We are also working on several areas to further improve TIARA. One area is the use of various application-specific features to build more accurate topic summarization results.We also plan to evaluate TIARA on other applications, such as visual patient records analysis, to help get more insight on how TIARA could help the user and which parts should be improved to make it more useful. Moreover, we are interested in recording users' usage logs and then extracting some interesting usage patterns that lead to the discovery of the insight.

## ACKNOWLEDGMENTS

We would like to thank Qiang Zhang, Lei Shi, and Xiaohua Sun for their help in the creation of TIARA. We would also like to thank Jonathan Feinberg and Martin Wattenberg for sharing the source-code of Wordle with us. Finally, we thank Fernanda Vïgas for providing Themail for our study, and J. P. Fasano and Matt Callcut for proofreading our article.

## REFERENCES

BIER, E. A., STONE, M. C., PIER, K. A., BUXTON, W., AND DEROSE, T. 1993. Toolglass and magic lenses: the see-through interface. In *Proceedings of the ACM SIGGRAPH Conference*. 73–80.

BLEI, D., NG, A., AND JORDAN, M. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res. 3,* 5, 993–1022.

BYRON, L. AND WATTENBERG, M. 2008. Stacked graphs - Geometry & aesthetics. *IEEE Trans. Vis. Comput. Graph. 14,* 6, 1245–1252.

CARENINI, G., NG, R., AND ZHOU, X. 2007. Summarizing email conversations with clue words. In *Proceedings of the International World Wide Web Conference (WWW)*. 91–100.

CHEN, Y., WANG, L., DONG, M., AND HUA, J. 2009. Exemplar-Based visualization of large document corpus. *IEEE Trans. Vis. Comput. Graph. 15,* 6, 1161–1168.

CLARKE, C. L. A., KOLLA, M., CORMACK, G. V., VECHTOMOVA, O., ASHKAN, A., BÜTTCHER, S., AND MACKINNON, I. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the ACM SIGIR Conference*. 659–666.

DON, A., ZHELEVA, E., GREGORY, M., TARKAN, S., AUVIL, L., CLEMENT, T., SHNEIDERMAN, B., AND PLAISANT, C. 2007. Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In *Proceedings of the Conference on Infornmation and Knowledge Management (CIKM'07)*. 213–222.

DREDZE, M., WALLACH, H., PULLER, D., AND PEREIRA, F. 2008. Generating summary keywords for emails using topics. In *Proceedings of the IUI Conference*. 199–206.

HAVRE, S., HETZLER, E., WHITNEY, P., AND NOWELL, L. 2002. Themeriver: visualizing thematic changes in large document collections. *IEEE Trans. Vis. Comput. Graph. 8,* 1, 9–20.

HE, X., CAI, D., AND NIYOGI, P. 2005. Laplacian score for feature selection. In *Proceedings of the NIPS Conference*.

HEARST, M. 1995. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'95)*. 59–66.

IWATA, T., YAMADA, T., AND UEDA, N. 2008. Probabilistic latent semantic visualization: Topic model for visualizing documents. In *Proceedings of the KDD Conference*. 363–371.

KERR, B. 2003. Thread arcs: An email thread visualization. In *Proceedings of the InfoVis'03 Conference*. 211–218.

LAN, M., TAN, C. L., LOW, H.-B., AND SUNG, S. Y. 2005. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *Proceedings of the WWW Conference* (Special Interest Tracks and Posters). 1032–1033.

LESKOVEC, J., BACKSTROM, L., AND KLEINBERG, J. 2009. Meme-Tracking and the dynamics of the news cycle. In *Proceedings of the KDD Conference*. 497–506.

LIU, S., ZHOU, M. X., PAN, S., QIAN, W., CAI, W., AND LIAN, X. 2009. Interactive, topic-based visual text summarization and analysis. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*. 543–552.

LUBOSCHIK, M., SCHUMANN, H., AND CORDS, H. 2008. Particle-Based labeling: Fast point-feature labeling without obscuring other visual features. *IEEE Trans. Vis. Comput. Graph. 14,* 6, 1237–1244.

MCCALLUM, A., WANG, X., AND CORRADA-EMMANUEL, A. 2007. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Intell. Res. 30*, 249–272.

MITRA, P., MURTHY, C. A., AND PAL, S. K. 2002. Unsupervised feature selection using feature similarity. *IEEE Trans. Patt. Anal. Mach. Intell. 24,* 3, 301–312.

NARDI, B., WHITTAKER, S., ISAACS, E., CREECH, M., JOHNSON, J., AND HAINSWORTH, J. 2002. Integrating communication and information through contactmap. *Comm. ACM 45,* 4, 89–95.

PERER, A. AND SMITH, M. 2006. Contrasting portraits of email practices: Visual approaches to reflection and analysis. In *Proceedings of the AVI Conference*. 389–395.

RENNISON, E. 1994. Galaxy of news: An approach to visualizing and understanding expansive news landscapees. In *Proceedings of the UIST'94 Conference*. 3–12.

SAHAMI, M. 1998. Using Machine Learning to Improve Information Access. Ph.D. thesis, Department of Computer Science, Stanford University.

SALTON, G. AND BUCKLEY, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag. 24,* 5, 513–523.

SARKAR, M. AND BROWN, M. H. 1994. Graphical fisheye views. *Commun. ACM 37,* 12, 73–84.

STASKO, J., GORG, C., AND LIU, Z. 2008. Jigsaw: Supporting investigative analysis through interactive visualization. *Inf. Vis. 7,* 2, 118–132.

STROBELT, H., OELKE, D., ROHRDANTZ, C., STOFFEL, A., KEIM, D. A., AND DEUSSEN, O. 2009. Document cards: A top trumps visualization for documents. *IEEE Trans. Vis. Comput. Graph. 15,* 6, 1145–1152.

VAN HAM, F., WATTENBERG, M., AND VIÉGAS, F. B. 2009. Mapping text with phrase nets. *IEEE Trans. Vis. Comput. Graph. 15,* 6, 1169–1176.

VENOLIA, G. AND NEUSTAEDTER, C. 2003. Understanding sequence and reply relationships within email conversations: a mixed-model visualization. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'03)*. 361–368.

VIEGAS, F., GOLDER, S., AND DONATH, J. 2006. Visualizing email content: Portraying relationships from conversational histories. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. 979–988.

WAN, S. AND MCKEOWN, K. 2004. Generating overview summaries of ongoing email thread discussions. In *Proceedings of the COLING Conference*. 549–555.

WANG, D., LI, T., ZHU, S., AND DING, C. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the SIGIR'08 Conference*. 307–314.

WATTENBERG, M. AND VIEGAS, F. 2008. The word tree, an interactive visual concordance. In *Proceedings of the InfoVis'08 Conference*. 1221–1228.

WISE, J., THOMAS, J., PENNOCK, K., LANTRIP, D., POTTIER, M., SCHUR, A., AND CROW, V. 1995. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the InfoVis'95 Conference*. 51–58.

Received July 2010; revised March 2011; accepted March 2011