Storylines: Visual Exploration and Analysis in Latent Semantic Spaces

Weizhong Zhu, Chaomei Chen

College of Information Science and Technology, Drexel University
3141 Chestnut Street
Philadelphia, PA, 19104

ABSTRACT

Tasks in visual analytics differ from typical information retrieval tasks in fundamental ways. A critical part of a visual analytics is to ask the right questions when dealing with a diverse collection of information. In this article, we introduce the design and application of an integrated exploratory visualization system called Storylines. Storylines provides a framework to enable analysts visually and systematically explore and study a body of unstructured text without prior knowledge of its thematic structure. The system innovatively integrates latent semantic indexing, natural language processing, and social network analysis. The contributions of the work include providing an intuitive and directly accessible representation of a latent semantic space derived from the text corpus, an integrated process for identifying salient lines of stories, and coordinated visualizations across a spectrum of perspectives in terms of people, locations, and events involved in each story line. The system is tested with the 2006 VAST contest data, in particular, the portion of news articles.

Keywords: latent semantic indexing, social network analysis

1 INTRODUCTION

Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces. Its goal is to detect the expected and discovery the unexpected [1]. The exploration and visualization of stories in news articles is a challenging task. The task could be characterized as four words, WHO, WHEN, WHERE and WHAT. Each word implies many questions. For instance, WHO may include questions such as who are the key players relevant to the story, how the relevant players are connected, which of the relevant players are deliberately engaged in what kind of activities and so on.

The basic idea for an effective data exploration is to include the human in the data exploration process and combine the flexibility, creativity, and general knowledge of the human with the intelligent support from text analysis algorithms. We firstly develop data signatures from the sources of the unstructured text and produce high-dimensional representations both statistically and semantically. The source may include news, email, citation, and web blog. The data signatures, such as single keyword, n-gram and named entity, are extracted directly from source corpus by natural language processing techniques. Then we try to discover the hidden, weak or unexpected relationships while considering the entire concept space.

Our system includes two major parts: story generation process and social network analysis. The key function of story generation is to visualize the entire dataset and provide an overview for exploratory analysis. Our design follows the well-known Information Seeking Mantra: overviews first, filter, and details on demand. The use of Latent Semantic Indexing (LSI) [2] gives users effective support for understanding the underlying information space. LSI dimensions are optimal solutions for the characteristic document vectors. From the Singular Value Decomposition (SVD) point of view, LSI dimensions are generally projection dimensions. According to the original paper of LSI [2], these dimensions are complex and can't be directly inferred. In this study, we propose a novel visual approach to explicitly represent the latent semantic dimensions in order to track the main topics in source data.

In order to identify key players, locations and organizations in stories, we combine co-occurrence analysis of named entities and importance measures such as degree centrality and betweenness centrality, which overcomes the weaknesses of pure entity frequency counting,

The rest of this paper is organized as follows: Section 2 reviews related work, while Section 3 focuses on tasks that Storylines will target. Section 4 reviews the system architecture, procedure and data pre-processing. Section 5 discusses key features of our system. Section 6 applies Storylines on VAST tasks. Section 7 presents discussion, conclusions and future work.

2 RELATED WORK

Latent Semantic Analysis (LSA) uses statistical machine learning in text analysis. SVD is a dimension reduction method. For a high dimensional dataset, SVD approximates the original semantic space with a much lower dimensionality, usually 100-400 dimensions.

Soboroff [3] used LSA to visually cluster documents based on the usage pattern of n-gram terms. Landauer [4] describes a linear SVD technique and applies it to a

collection of a half billion documents containing 750 000 unique word types. LSA presumes that the overall semantic content of a document can be approximated as a sum of the meaning of its words. According to Landauer's study, an effective LSA representation of documents must start by deriving a good high dimensional semantic space for the entire domain. Researchers have found that the optimally reduced dimensionality to match human interpretation is about 300 dimensions. One argument made by Landauer et al. is that there is no guarantee that any particular projection will reveal something familiar or new to the human visual system. They used the GGobi [5] high dimensional data viewer to display any subset of three dimensions. GGobi can automatically find dimension triplets and rotations to maximize properties such as dispersion or grouping of points. They generated visualizations by user-guided projection-pursuit methods, which is an interactive process of examining subspace views and marking data points of interest.

Ding [6] developed a probabilistic model for latent semantic indexing based on the dual relationship between words and documents. According to this model, LSA is the optimal solution of the model. The model has established the amount of contributions of dimensions to the latent semantic space. Specifically, the singular value squared is the amount of contributions of the corresponding dimension. This quadratic dependence finding indicates that LSA dimensions with small singular values are overrepresented by the linear relationship as previously thought. Furthermore, they demonstrated that the importance of LSA dimensions follows the Zipf-distribution, which means that a small number of most important dimensions can adequately approximate the overall semantic space.

The visual analytics literature per se is relatively small but rapidly growing. For example, Wilkinson et al. [7] describe visual analytic techniques for high-dimensional multivariate exploratory data analysis using guided pairwise views of point distributions. Wong et al. [8] present a visual analytics technique to explore graphs using the notion of graph signature. Their work shows the advantages and strengths of graph signature-guided approach over traditional graph visualization approaches. Shen et al. [9] develop a visual analytics tool, OntoVis, for understanding large, heterogeneous social networks, in which nodes and links could represent different concepts and relations. These techniques provide an encouraging context for our work. Visually and systematically exploring latent semantic spaces share some common challenges with these advanced techniques. On the other hand, our approach is unique with its mathematical basis of LSI and innovative integrations of social network analysis in understanding salient dimensions involved in unstructured text.

3 TASKS FOR VISUAL ANALYTICS

Our aim is to support the following visual analytics tasks:
Obtain an overall visualization of the semantic space defined by a collection of text documents
Understand the dimensionality of the semantic space
Identify the role of a specific dimension in terms of a space of terms
Identify the nature of a specific dimension in terms of most contributing terms and identify documents of interest
Explore the document space and generate storylines
Explore a storyline by visually examining networks of named entities and their inter and intro relationships extracted from the storyline

4 SYSTEM ARCHITECTURE

Storylines includes four major components: natural language processing, LSI based feature selection and topic detection, timeline information filtering and social network analysis.

Latent semantic indexing (LSI) is the foundation of the work. The design goal is to provide an LSI-driven and visually accessible user interface to the latent semantic space. LSI dimensions are used to organize the interactive visualizations of the semantic space systematically. The most significant contribution of the work is that it not only visualizes previously latent semantic spaces but also provides analysts with direct access to individual dimensions so that they can systematically explore and study the significance of a dimension with references to terms and documents.

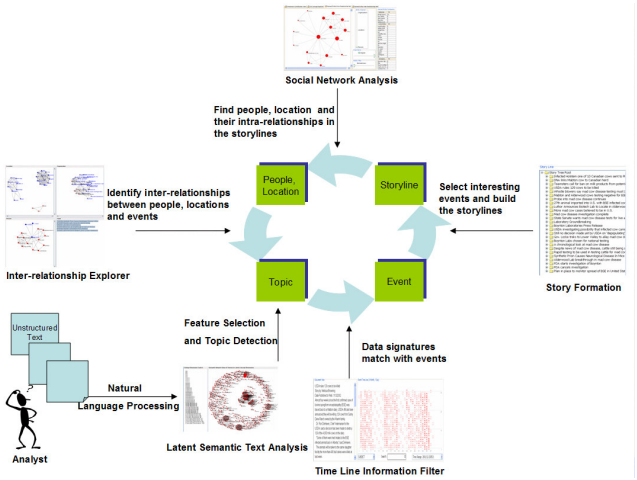

Figure 1 The Architecture of Storylines.

4.1 Procedure

Motivated by the above research result, we construct the latent semantic space as a semantic network as follows:
Step 1: Use SVD to decompose the term-document matrix into term-concept matrix and document-concept matrix. For VAST dataset, it converges in 1 178 dimensions, see more details in the section of data pre-processing.

Step 2: In the term-concept matrix, the top K dimensions are selected to explore the themes and topics. The contribution of every dimension to the latent concept space is showed in Fig. 2. It showed the top dimensions make the most contribution to the latent concept space. We select 100 dimensions. Based on the observation, each dimension could be represented by a list of distinct themes and could include one or more topics. In a dimension k, theme selection is based on the threshold on x_{ik}^2, where x_{ik} is the term X_i projection score in the k^{th} dimension. x_{ik}^2 could be thought as the local contribution of term X_i to the k^{th} dimension. The threshold is experiment-driven and is varied from datasets. The theme selection starts from the dimension with a higher singular value score and obeys a step-wise strategy. That means if a theme is selected in a dimension with a higher singular value, it won't be counted in later dimensions, in case that a theme appears in many dimensions.

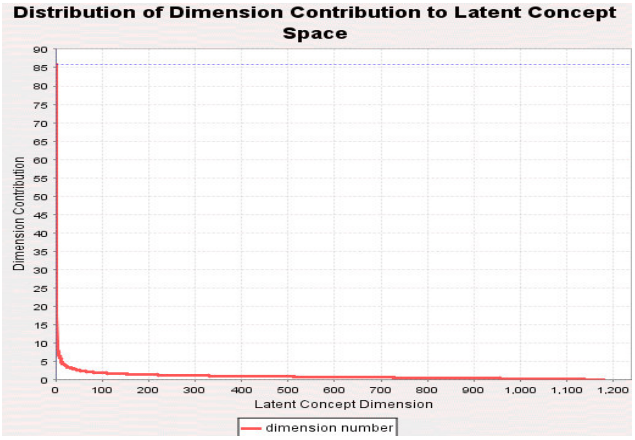

Figure 2 The distribution of dimension contribution to the latent concept space. Dimension contribution is measured by S^2, where S is the singular value of the dimension.

Step 3: For a dimension k, the contribution W_{ik} of a term X_i is calculated by $S^2 x_{ik}^2$, where S is the singular value of the dimension k and x_{ik} is the k^{th} dimension projection score. W_{ik} could be thought as the contribution of term X_i to the dimension k in a global sense. An overall contribution of a term X_i to the latent concept space is

$$\sum_{n=1}^{m} w_{in} \qquad (1)$$

where m is the number of selected dimensions that represents the whole latent concept space. The SVD method has the disadvantage of losing the interpretability of features if only using dimensions or dimension projection scores. Using features with the highest sum score of W_{ik} derived from the probabilistic LSI model [6] overcomes the problem and can be interpreted as the most important features for the latent concept space. If use percentage, the contribution of a term is

$$\sum_{n=1}^{m} w_{in} \Big/ \sum_{i=1}^{t} \sum_{n=1}^{m} w_{in} \qquad (2)$$

where t is total number of terms in the corpus. Term selection algorithm mentioned in step 2 iteratively picks themes until these themes collectively make 80% contributions to the entire latent concept space based on equation (2). 979 distinct themes are selected from the top 100 dimensions for the VAST data set.

Step 4: In top K dimensions, the contribution distribution of a term X_i is represented by a vector of W_{ij}, where j belongs to {1, k}. The cosine similarity between the contribution vectors of two distinct terms weights their distance in the latent semantic concept. The coordinates of these terms in the latent concept space are assigned according to Kamada Kawai graph layout algorithm [10].

Step 5: Using the same algorithm, the latent concept space could be represented as a semantic document network. In the document-concept matrix derived from SVD, each document is represented as a vector of document contribution distribution to the selected top concept dimensions.

4.2 DATA

The 2006 VAST contest data contains fictitious information. It was created for testing and evaluation of visual analytic tools only. It consists of text, telephone logs, photos, and other data. In this study, we focus on the text document collection, which contains 1182 synthesized documents.

The key question to be answered for the original contest was: What is the situation in the scenario of a fictitious small town Alderwood [11] and what is your assessment of the situations? The original VAST contest included the following visual analytic tasks:

1. Who are the players relevant to the plot?
2. Which of the relevant players are innocent bystanders?
3. Which of the relevant players are deliberately engaged in deceptive activities?
4. How are the relevant players connected?
5. What is the time frame in which this situation unfolded?
6. What events occurring during this time frame are relevant to the plot?
7. What locations were relevant to the plot?
8. What, if any, connections are there between relevant locations?
9. What activities were going on in this time frame?
10. Which players are involved in the different activities?

Storylines is designed to provide a framework that can facilitate systematic investigations of these questions from the perspective of visual exploration and analysis of a latent semantic space formed by the collection of unstructured text. Some of the questions are more appropriate to be addressed with multiple sources and types of data. Nevertheless, as the first step towards a text-centric visual analytics tool, we will focus on properties of such latent

semantic spaces and operations that could help us better understand the structure of a latent semantic space.

The complexity of understanding a latent semantic space as a whole and the need to analyze the role of individual dimensions of the space intuitively is a significant challenge for text analysis based on the concept of latent semantic indexing. Storylines demonstrates some innovative ideas to improve the understanding of a usually rather evasive notion of a latent semantic space.

4.3 Data Pre-processing

Data pre-processing directly influences the quality of the visualization. First, we convert the 1 182 news articles in the VAST contest dataset to a more structured XML format. Next, stopword filtering, STANFORD part-of-speech (POS) tagging [12], Port stemming and n-gram selection [13] are applied to the corpus. A total of 13 343 single noun keyword, noun bi-gram, noun tri-gram and noun quadra-gram are selected for subsequent text analysis. We define a heuristic rule to select noun n-grams. Words in each noun n-gram must be labeled as either all NN (noun) or composite JJ (adjective) and NN by the POS tagging. Based on the previous latent semantic indexing analysis, noun n-grams are added to generate the term-document matrix. The associative relationship between a term and a document is weighted by traditional TFIDF. We assign additional weights to n-grams. For instance, bi-gram gains extra weight by timing 2 and tri-gram times 3. We believe n-grams will add more semantic meanings to the latent concept space. The corresponded latent conceptual space should be more structural and organized because these n-grams bring in constraints such as hieratical relationships or associative relationships between n-grams and single key words. The named entities, person names, locations, organizations and male or female pro-noun, are extracted with Lingpipe [14]. In this system, if a named entity has a same co-inference id as that of another name entity, the two named entities will be unified as one.

5 KEY FEATURES

Storylines has two parts in terms of its functionality: generating a visually accessible latent semantic space, and analyzing thematic threads in the form of story lines.

5.1 Generating a Visually Accessible Semantic Space

A major barrier between analysts and a large complex corpus of unstructured text is the complexity of the underlying structure. Traditionally, the unit of analysis tends to focus on terms and documents. Although works such as LSI provide a useful concept-based perspective, the effectiveness of access is usually undermined by its intrinsic complexity.

5.1.1 Visualizing an LSI Latent Semantic Space

The frequency of latent dimensions identified by LSI follows the Zipf-distribution [4]. Dimensions with smaller contributions can be safely omitted. For instance, in the VAST data set, dimension 1 makes 46 times more contribution than dimension 100, (S^2 values, 85.93 versus 1.83). In fact, adding more dimensions may reduce the accuracy of the latent semantic model. Instead of building a commonly seen co-occurrence network of terms, we compute a semantic network of terms based on their contributions to the most salient concept dimensions so that groupings of terms can be used as surrogates of latent dimensions. We choose the top 100 latent dimensions identified by LSI to generate term vectors. Each term is represented by a 100-dimension vector. The vector coefficients are the corresponding strengths of term projections on these dimensions. The similarity between two terms is defined by the cosine dot product of the two vectors. In the top 100 dimensions, our term selection algorithm picks 979 terms, which explain approximately 80% of the latent semantic space. The rest of 12 364 terms explained the remaining 20%. In terms of the importance, the content of the 80% coverage is statistically more important content than that of the 20%. It suggests the term contribution in top dimensions follows Zipf's law.

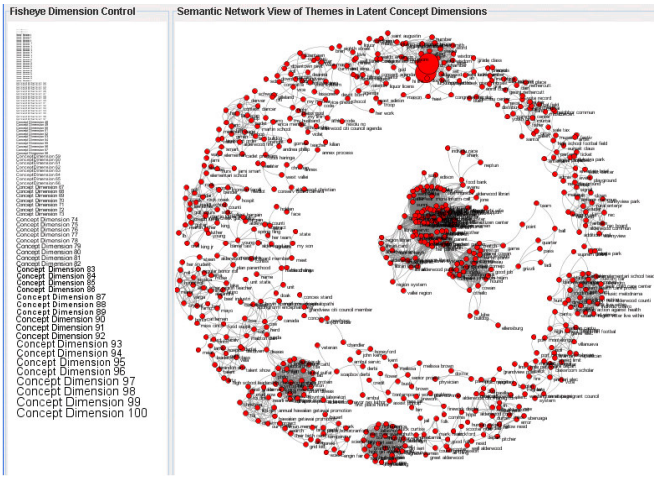

Figure 3 An overview of a semantic network of terms based on the top 100 most significant latent semantic dimensions in the VAST data set. The list of concept dimensions is filtered through a fisheye view for easy selection. Each term is represented by a red node. The node size reflects the global contribution of the term to the whole latent concept space.

Fig. 3 shows an overview of a semantic network of terms generated by this method. Each node represents a term. Links between nodes represent how similar they are in terms of their contributions to the top 100 latent concept dimensions. Topics associated with the clusters in the semantic network can be easily identified, including "Alderwood daily lucky number", "council events", "school events", "sports events", and "mad cow disease".

The contribution profile of a term is shown as the distribution of the contributions of the term across the top

100 dimensions, which is simply a display of the corresponding term vector. Right clicking on a node in the network will bring up a pop-up menu and from the pop-up menu one can select "Contribution Profile."

Fig. 4 depicts a chart that shows the contribution distribution of a term "mad cow disease" across the top 100 dimensions. It shows that this term has the strongest presence in the 4th and 5th dimensions. This feature can help analysts move from a single term to the most relevant latent concept dimension. Dimensions identified in this way can be used for query expansion, for retrieving relevant documents based on their projects on these dimensions, and for differentiating the appearances of the same terms in different contexts. Analysts could control the dimensions and seek the most related themes in each dimension from the fisheye menu (See Fig. 5).

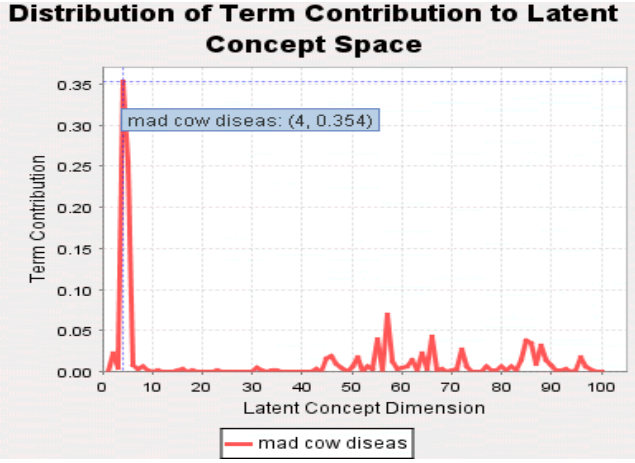

Figure 4 Identifying dimensions in which the term "mad cow disease" has a strong presence from a network of terms derived from the top 100 most representative latent concept dimensions. In this example, the term has the strongest presence in the 4th dimension.

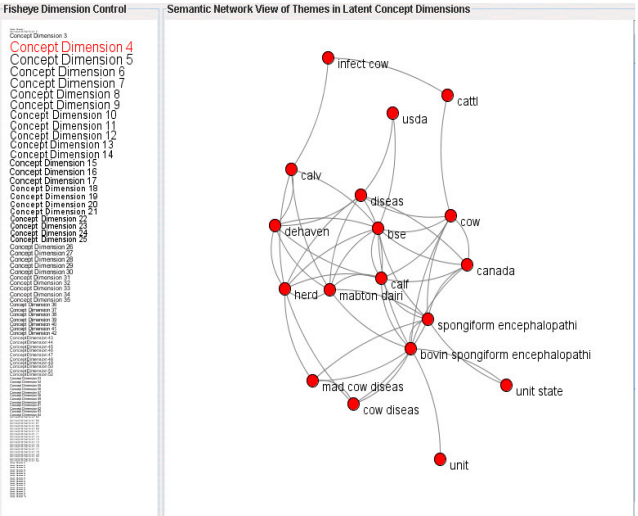

Figure 5 A semantic network of terms revealing the nature of the 4th dimension identified by LSI in the VAST data set.

Characterizing a latent concept dimension in terms of an associative network of contributing terms can help analysts in a sense making process of visual analytics. A network view of a single dimension provides a useful alternative way to capture the nature of a dimension, which has been a challenging task in dealing with latent semantic spaces. A network view reveals the intrinsic structure of a dimension. When combining a number of dimensions, the corresponding network view represents the intrinsic structure of a subspace of the latent semantic space. Unlike the commonly used term list to summarize the nature of a latent dimension, networks are more flexible and comprehensive. Networks preserve the term-term relationships that are unique to the underlying dimension or a subspace. Therefore, the semantics of the networks is clearly defined.

5.1.2 Story Formation

Storylines generates clusters of documents based on their keyword matching in the document space. Such clusters of documents form a storyline. Storyline starts with an overview of the entire collection of documents so that analysts can explore the distribution of documents over time by key word search. These key words are suggested by the latent concept exploration. Then analysts select the relevant documents and put in a time-order tree to generate stories.

Figure 6 shows the GUI of Storyline. It is divided into two levels. The top level view corresponds to the document level for interactively reviewing documents and incrementally adding key words to pick documents. Three coordinated views of 1,182 news articles in the VAST dataset are generated. Each red square in the Timeline View represents a news article. These articles are arranged as a monthly calendar. A column corresponds to a month and rows correspond to dates. The subject line of an article will pop up at the top of the view when the mouse cursor is over its square. Both subject-line search and full-text search are supported. The bottom level view corresponds to storylines, i.e. thematic clusters of documents.

Our system firstly gives users an overview of all the news articles in the corpus. The design follows a book metaphor and includes a document view, a document timeline, and an indexed term view. In the document view, details of each document become available upon double-clicking the red square representation of the document. The document timeline arranges documents along monthly columns. User can use either keyword filter or time filter to seek documents. The term view displays a list of all index terms ranked by their TFIDF weights.

The semantic network view described earlier provides an interactive interface to the latent semantic space identified by LSI. The single concept dimension view retrieves the

most associated terms from the term-concept space of LSI and suggests analysts possible themes for further searching. This is one of the biggest advantages of our system.

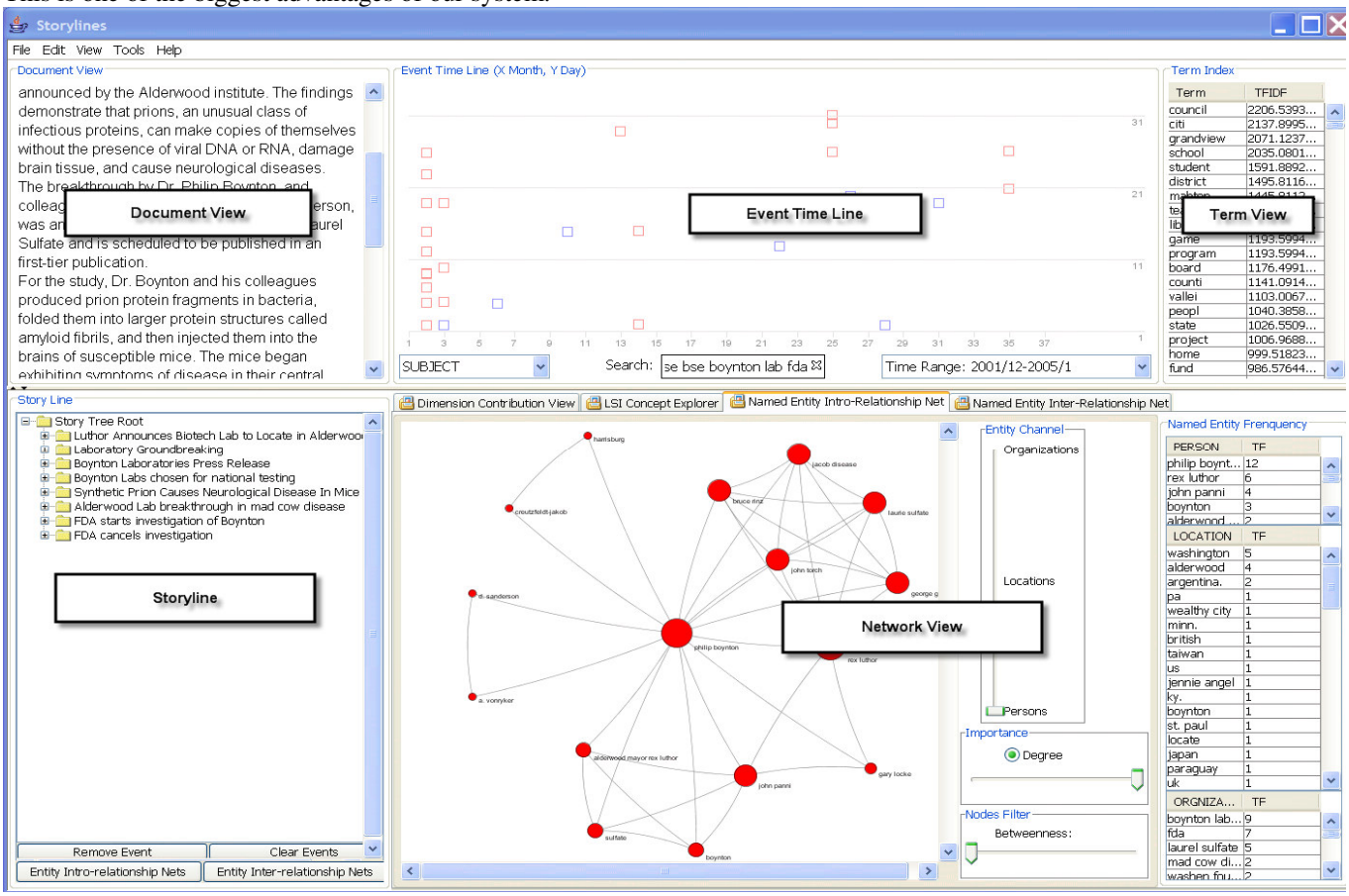

Figure 6 Story generation. A story is generated when a document is selected from the Timeline view. The red square of the document will turn in to a blue square. The generated story will appear in the Story Line window at the lower left corner of the interface. Each storyline is represented as a tree structure, containing subject line of document, time and named entities extracted from the corresponding documents. This feature allows analysts to study people who are involved in a particular thread of events as well as the locations and organizations related to these events.

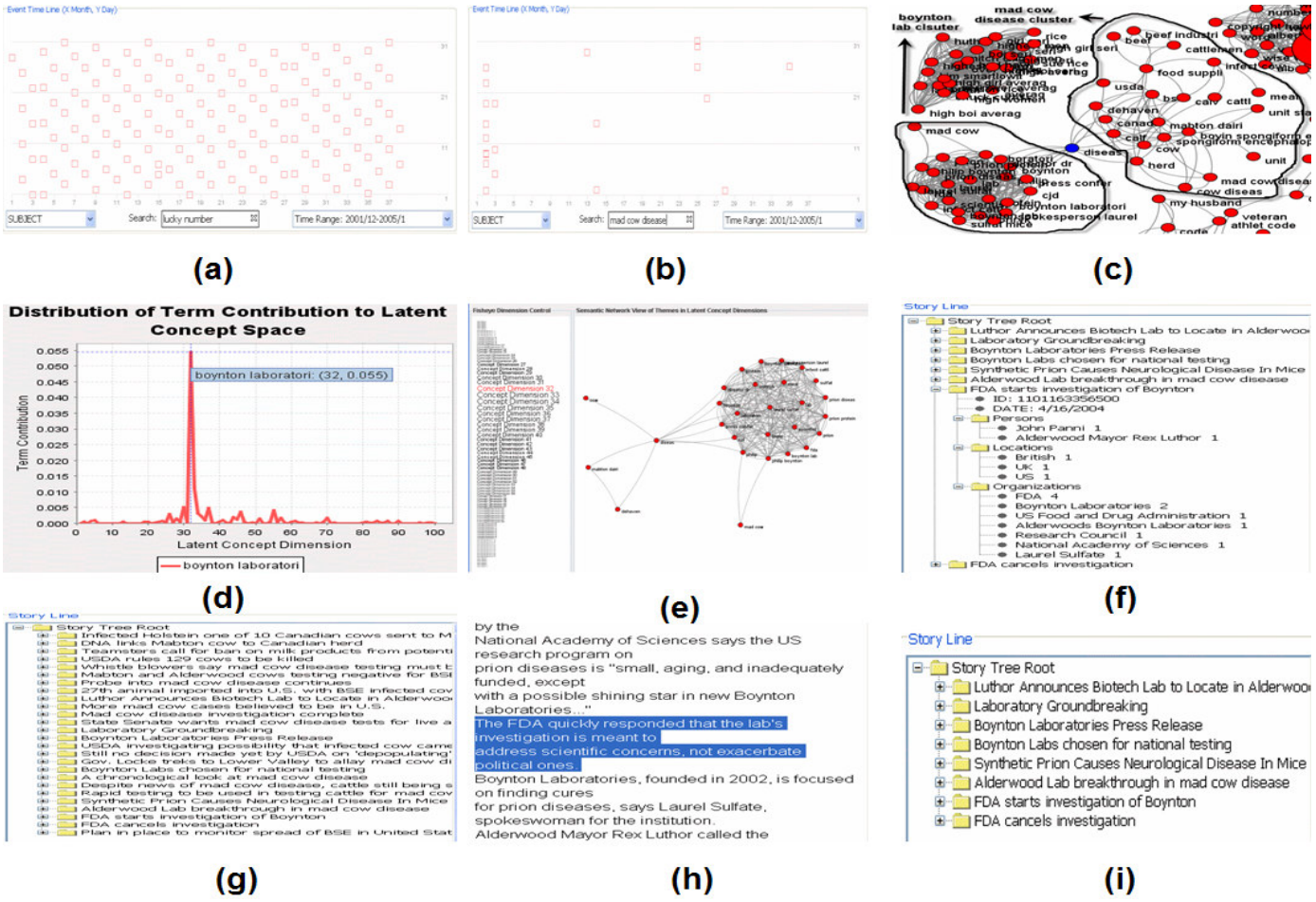

Figure 7 Story Generation Procedures.

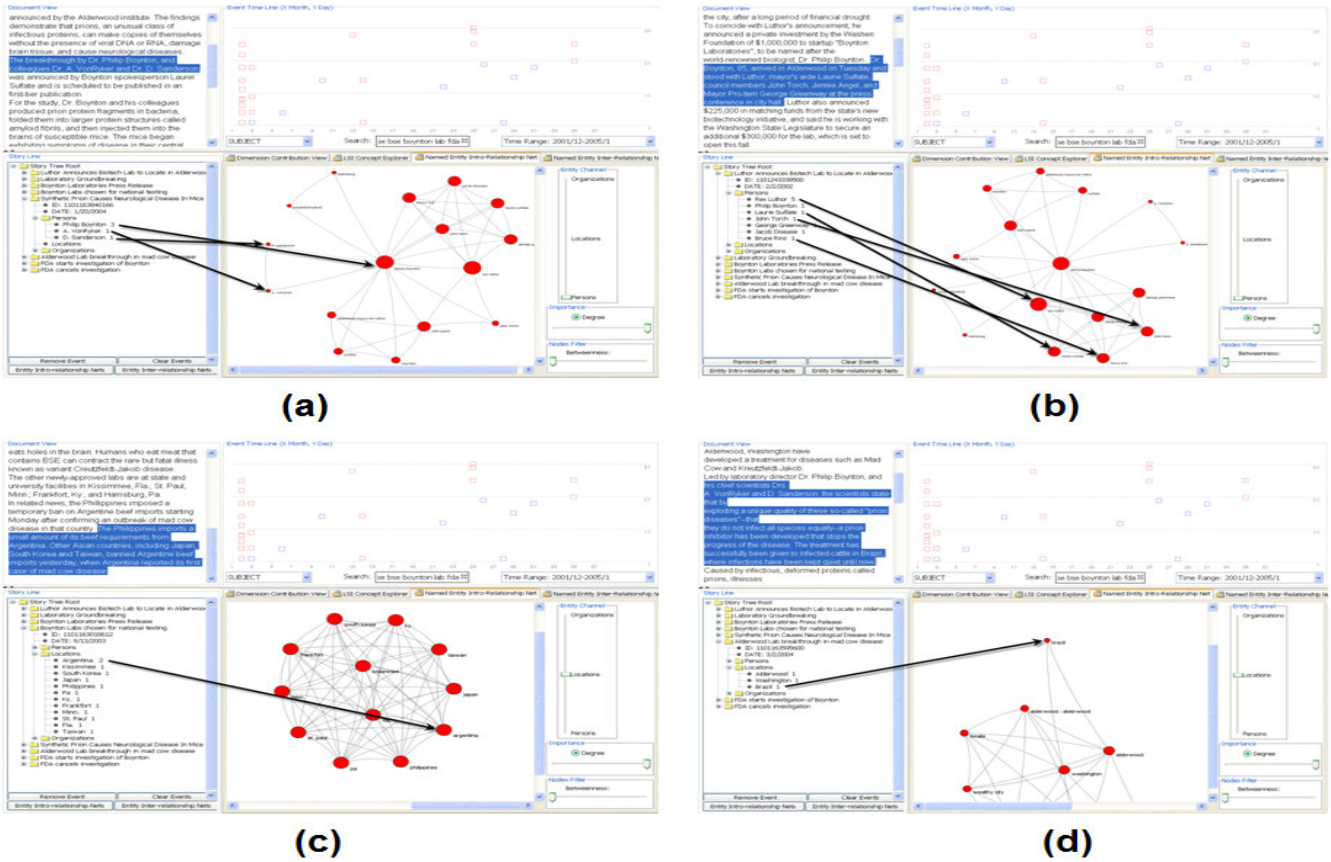

Figure 8 Social Network Analysis Procedures.

Users can interactively review documents, incrementally add key words, and pick documents. Then users can select the important documents that they thought could generate a story line and throw in an event tree, see Fig. 6. This feature provides users more flexibility to make stories.

5.2 Identification of Key Players and Locations

Another advantage of Storylines is its ability to identify key players in a social network according to their centrality measures. Counting terms and named entities such as persons, locations, and organizations is not particularly effective in finding influential members hidden in a group or a society because influential members may not necessarily have a high profile in terms of their appearances in news articles. Instead of focusing on frequency-based ranking mechanisms, we compare the topological properties of their connectivity in a network, namely using degree and betweenness centrality measures [15], which is a common method for social network analysis, but rarely used in semantic networks. Using centrality-based measures means that Storylines can identify people even when they do not have a prominent frequency profile but are important in keeping the network together. For example, in the "Boynton Lab" storyline identified in this study, the suspects' names did not have a high frequency-appearance in the corpus, but it is apparent from the depiction of the network that they appear to be members of a small group connecting to the key player of the story; therefore, they were potentially important players in this storyline, see named entity network view in Fig. 6. It is a named entity network of persons derived from a chosen storyline on FDA's investigation. A summary view on the right lists persons, locations, and organizations identified in news articles of this storyline.

6 Task Analysis

6.1 Result and Discussion

We investigate plots related to mad cow disease as an example to assess our system. Three plots are relevant to mad cow disease. One is the investigation of mad cow disease, "where did these mad cows come from". Another one investigates how mad cow disease affects the beef industry of Alderwood. The third focuses on a national lab called Boynton Biotech Lab built for mad cow disease testing and relevant biological research. Because our analysis showed FDA investigated the Boynton lab to address some kind of scientific and political issues, we decided to explore this plot.

6.1.1 Feather Selection and topic detection

The clusters in the latent concept space showed there are several major topics in the corpus of news, such as daily lucky number, council events, school events, events related to mad cow disease, sports events, Obituaries, and political events if you explore the top 10 dimensions. We firstly explored the top three topics, daily lucky number (see Fig. 7a), council school events and sports events. The distribution of these events in the Event Time Line shows that they are common events to the local community and are unlikely to be associated with unexplained underlying activities. In contrast, the distribution of the events related to mad cow disease is not even (see Fig. 7b), suggesting these documents may be connected to a number of potential events of interest. So we decided to explore the topic "mad cow disease" in more detail. Then we use latent space explorer to seek further evidence. From Fig. 3, a concept cluster related to "mad cow disease" could be easily identified. Moreover, a concept cluster "Boynton Laboratory" is linked with the one of "mad cow disease" (See Fig. 7c). Figure 7d shows that the term "Boynton laboratory" is particularly essential to the 32nd latent concept dimension with a highest contribution across the top 100 LSI dimensions. In order to explore the story hidden in the two linked theme clusters, we generated a combined key words list, "mad cow disease, bse" in dimension 5 and "Boynton laboratory, fda" from dimension 32 (see Fig. 7e) for collecting all the relevant documents.

6.1.2 Story Line Formation

The story generation process collects the retrieved documents selected in the Event Time Line by right double clicking each red square. If the document is added to the Story Line, the color of the square will turn into blue. Story Line organizes the events in a tree structure, see Fig. 7f. A top-level tree node represents a document with its subject line as the heading. Each document node has five children nodes, namely ID, Date, Persons, Locations and Organizations. Persons, locations, and organizations are named entities extracted from the original text.

Storylines allows users to form story lines that correspond to one or more clusters of documents. The resulting storyline is showed in Fig. 8g that shows a combined story line about *Mad Cow Disease* and *Boynton Lab*. The events are related to three threads of interest: 1) the origin of the mad cows 2) impacts on beef industry in the Alderwood community, and 3) a local research lab doing mad cow disease testing and related biological research. One of the news articles in the storyline list is entitled "FDA starts investigation of Boynton", which appears to be relevant for our analytic tasks (See Fig. 7h). A document in the Boynton Lab threads an FDA investigation. The nature of the investigation is to address "scientific concerns".
We decided to check this article in detail. According to the highlighted lines in the document, FDA investigates some scientific issues.

It is clear from the order of time view in Fig. 7i that FDA began the investigation right after the lab announced a research breakthrough. We hypothesized that certain members of this lab may have done something suspicious that triggered the FDA investigation. So we decide to keep events directly related to Boynton lab in the storyline and filter out all other events by the "Remove Event" function. Then we constructed a social network based on these events for further analysis. As a result, we generated a Boynton timeline with 8 events (See Fig. 7i).

As shown in Figure 7c, two clusters are marked: *Boynton Lab* and *Mad Cow Disease*. As explained earlier, the term-contribution similarity network captures latent concepts. Each latent concept primarily corresponds to a latent concept dimension. Thus we can see the two clusters as two latent concepts. The meaning of each of the concepts is represented by the component nodes and their interrelationships. For example, the *Mad Cow Disease* cluster contains terms such as "bse, calf, herd, and diary", whereas the *Boynton Lab* cluster contains terms such as "fda, prion, mad cow and protein".

Storylines has revealed that the Mad Cow Disease cluster corresponds to the 4th LSI dimension, whereas the Boynton Lab cluster corresponds to the 32nd LSI dimension. It is usually not easy to tell which clusters are more prominent in the latent semantic space identified by LSI. However, since Storylines arranges the dimensions in the order of their contributions, we know that the 4th dimension is more important in the semantic space than the 32nd dimension; therefore, the *Mad Cow Disease* concept is more important than the *Boynton Lab* one.

6.1.3 Social Network Analysis

Clicking on the "Entity Intro-relationship Nets" button generates a network of named entities. The named entity network includes three channels, namely, Person, Location and Organization. The size of a node reflects the relative importance measured by its degree centrality in the network. Named entity summary view shows the frequency rankings of these named entities.

Storylines makes it easy to study social networks associated with a given storyline. For example, the analyst wants to find not only all the names mentioned in the storyline documents, but also how they are connected to one another. If we are interested in pursuing the FDA investigation thread, we would be interested in people who were involved in the investigation. First, we consider everyone appeared in the network as suspects in the FDA investigation. We could easily read through the eight documents of the Boynton storyline in Storylines. Evidence

showed that suspects could be narrowed down to two clusters. As to be described in next section, one is a group of people involved in a scientific discovery and the other a group of people who have political ties (See the named entity network view in Fig. 6. 8a-b). Figure 8a focuses on a 'scientific discovery' cluster in the named entity network of people in the Boynton storyline. The three people pointed in the figure were involved in a scientific discovery. The original news article shown in the Document View indicates that they were members of the Boynton lab. Figure 8b explores another cluster in the named entity network of people in the Boynton storyline – an outsider cluster. The original news article shown in the Document View highlights the activity involved by this group of people, namely the mayor, and council members at an event related to the startup of the Boynton lab (See Fig. 8b). Unlike the small 'scientific discovery' group, most members of the second cluster are not the members of the Boynton lab. But there is one exception that the previous council secretary shifted to Boynton lab and became the spokeswoman. The social network intuitively reveals the connections between the two major groups in the Boynton storyline. People in these clusters could be treated as suspects who have deceptive activities. Actually the names of these people appear in the answer sheet of VAST tasks.

6.1.4 Identification of Locations

Locations of the plot are identified from the named entity network of locations (See Figures 8c-d). Locations "Argentina" and "Brazil" are related to Boynton storyline. The content of the document illustrates the Boynton lab tested their experiments on infected cow in Brazil.

6.1.5 Identification of Named Entity Inter-relationships

Fig. 8a-d show the investigation on the intra-relationships within the same type of named entity. An entity inter-relationship explorer (see Fig. 9) supports interactive and dynamical exploration of concurrence associations between single nodes or the clusters in the named entity networks and the context of events in a storyline, such as subject line, time etc. Then further hypothesis and investigations are easily formed and performed by analysts that aware contextual information of the whole story. Clicking on the "Entity Inter-relationship Nets" button triggers the explorer.

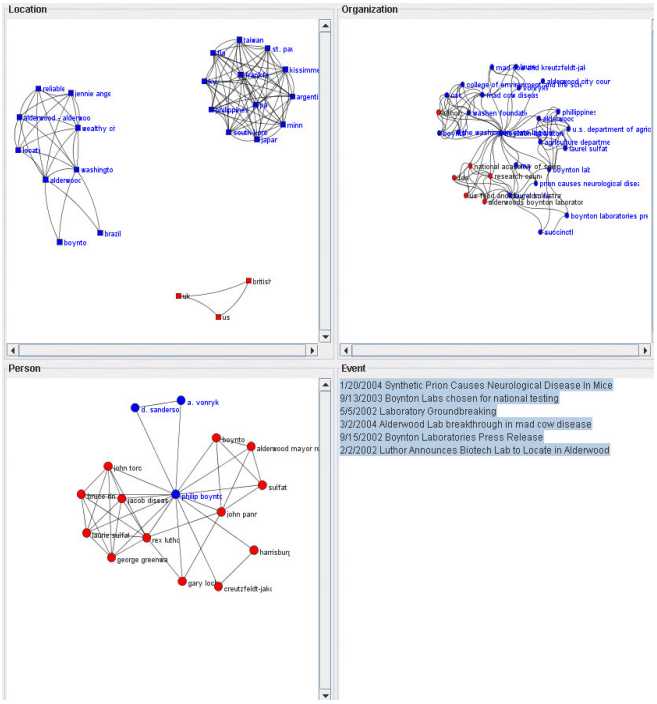

Figure 9 Named Entity Inter-relationship Explorer. After selecting a cluster in one type of named entity associative network, for instance people, related location entities, organization entities, subjects and time of events are highlighted accordingly. The color of nodes and labels is changed from red to blue.

7 Discussions and Conclusion

7.1 Contributions to Visual Analytics

We have made several observations of the potential of Storylines. First, its primary novelty is the support for explicit and direct visual exploration of latent semantic spaces identified by LSI. This novelty is potentially extensible to other models of text, such as generative models in general. Second, the work has made the first step towards an integration of text analysis and social network analysis and using network visualization to facilitate sense making processes involving high-complexity and high-dimensionality problems. In comparison to alternative approaches such as virtual tours through the entire latent spaces, our approach has several advantages: 1) it supports systematic exploration of the data with reference to quantitative measures of importance, i.e. singular value squared, 2) networks of terms by their contributions to the underlying dimensions provide a unique way to understand the nature of a dimension and differentiate different dimensions, and 3) analysts have a clear idea of the extent their visual exploration covers the latent space. Third, we emphasize the role of association in reasoning and investigations. Operations are triggered by association

whenever possible. Although we do provide users with search functions, the use scenario primarily focuses on exploratory analysis and supporting the level of flexibility required by such analytic processes.

7.2 Future Work

This is the first step of an ongoing research program. The ultimate goal is to reduce the complexity of analyzing a latent semantic space of unstructured text to the level of exploring a well structured body of text. There are many unsolved issues. A number of more specific challenges need to be addressed in the future work. For example, an optimal approach to select the dimensionality of the subspace of the latent concept space would be based on the optimal number of dimensions that peak the dual probabilistic model in which LSI is the optimal solution. This involves estimating the optimal number of dimensions and it will increase the ease of use without imposing a threshold in advance. Another example of improvement would be more coordinated and tightly coupled visual analytics features across all levels, namely, the dimension level, the term level, the document level, and the storyline level. As shown in the visualized networks, not all terms belong to clearly bounded clusters. The boundaries between terms could be less distinct than the prominent clusters such as the *Boynton Lab* and *Mad Cow Disease*. It would be a useful feature if one can add additional dimension-specific overlays on top of a given network. This feature will help analysts to identify concepts that do not lend themselves to visually salient clusters. For example, upon selecting a specific dimension in the context of a given subspace, the interface could highlight the terms that belong to the selected dimension.

The VAST contest data is a synthesized dataset. Our next step is to extend the work to real-world datasets such as news archives, live news feeds, email archives, citation records and web blogs. A thorough task analysis is in order for a better understanding of an optimal task-oriented design to support visual analysis of unstructured text. Future work should also involve multimedia data.

In conclusion, Storylines represents a new way to visually explore and systematically study a latent semantic space derived from unstructured text. It provides novel features to facilitate analysts to identify plausible thematic threads with no assumption of prior knowledge of the subject domain. The integration of text analysis and social network analysis has demonstrated its values in sense making processes of visual analytics. The innovative integration of visualization and latent semantic indexing has the potential to make wider impacts on text analysis as well as visual analytics.

Acknowledgements

The work is in part supported by the National Visualization and Analytics Center (NVAC) through the Northeast Visualization and Analytics Center (NEVAC) and the National Science Foundation under Grant No. SEIII-0612129. The authors would like to thanks the VAST contest organizers for making the dataset available.

REFERENCES

[1] James J. Thomas and Kristin A. Cook. Illuminating the Path: The research and development agenda for visual analytics. National Visualization and Analytics Center; 2005

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. Indexing by Latent Semantic Analysis. Journal of the Society for Information Science 1990; 41(6):391-407.

[3] Soboroff, I. M., Nicholas, C. K., Kukla, J. M., and Ebert, D. S.. Visualizing document authorship using n-grams and latent semantic indexing. Proceedings of the 1997 Workshop on New Paradigms in information Visualization and Manipulation. New York, NY; 1997, p. 43-48.

[4] Thomas K. Landauer , Darrell Laham, and Marcia Derr. From paragraph to graph: Latent semantic analysis for information visualization. PNAS, April 6, 2004; 101(Suppl. 1):5214-5219.

[5] http://www.ggobi.org/

[6] Ding, C. H.. A probabilistic model for Latent Semantic Indexing. Journal of the Society for Information Science 2005; 56(6):597-608.

[7] Leland Wilkinson, Anushka Anand, and Robert Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. IEEE Transactions on Visualization and Computer Graphics 2006; 12(6):1363-1372.

[8] Pak Chung Wong, Harlan Foote, George Chin Jr., Patrick Mackey, and Ken Perrine. Graph signatures for visual analytics. IEEE Transactions on Visualization and Computer Graphics 2006; 12(6):1399-1413.

[9] Zeqian Shen, Kwan-Liu Ma, and Tina Eliassi-Rad. Visual analysis of large heterogeneous social networks by semantic and structural abstraction. IEEE Transactions on Visualization and Computer Graphics 2006; 12(6):1427-1439.

[10] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. Information Processing Letters 1989; 31:7-15.

[11] http://www.cs.umd.edu/hcil/VASTcontest06/dataset.htm

[12] http://nlp.stanford.edu/software/tagger.shtml

[13] LeeFeng Chien, T. I. Huang, M. C. Chien. Pat-tree-based Keyword Extraction for Chinese Information Retrieval. Proceedings of SIGIR 1997; 1997, p. 50-58.

[14] http://www.alias-i.com/lingpipe/

[15] L. Freeman. Centrality in social networks: Conceptual clarification. Social Networks 1979; 1:215–239.

Weizhong Zhu Could be contacted at wz32@drexel.edu
Tel: +1 (215) 895-6627
Fax: +1 (215) 895-2494

Chaomei Chen Could be contacted at
chaomei.chen@ischool.drexel.edu
Tel: +1 (215) 895-6627
Fax: +1 (215) 895-2494