SentiView: Sentiment Analysis and Visualization for Internet Popular Topics

Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang

***Abstract*—There would be value to several domains in discovering and visualizing sentiments in online posts. This paper presents SentiView, an interactive visualization system that aims to analyze public sentiments for popular topics on the Internet. SentiView combines uncertainty modeling and model-driven adjustment. By searching and correlating frequent words in text data, it mines and models the changes of the sentiment on public topics. In addition, using a time-varying helix together with an attribute astrolabe to represent sentiments, it can visualize the changes of multiple attributes and relationships among demographics of interest and the sentiments of participants on popular topics. The relationships of interest among different participants are presented in a relationship map. Using a new evolution model that is based on cellular automata, it is able to compare the time-varying features for sentiment-driven forums on both simulated and real data. Adaptable for different social networking platforms, such as Twitter, blog and forum, the methods demonstrate the effectiveness of SentiView in analyzing and visualizing public sentiments on the Web.**

***Index Terms*—Microblog, sentiment, social networks, visual analytics, web forums.**

I. Introduction

With the rapid development of Web technologies, an increasing number of social networking platforms have been widely used, such as blogs, forums, and microblogs. They have become indispensible for public information sharing. As one of the most active forums in China, Tianya Community Forum,[1] has more than 45 million registered users, over 500 thousand daily users on average,[2] and tens of thousands of new topics posted every day. There are 503 million of subscribers and more than hundred posts per day in the Sina microblog, one of the most popular microblogs in China. Thus, every day, millions of net surfers view and comment topics on Web forums. The information in such forums is multidimensional, time varying, and mutable, so it is difficult to analyze and visualize the features of popular topics.

This paper presents an interactive visualization system, SentiView, which analyzes public sentiments from text posted via media such as forums and predicts the short-term trend of the sentiments about events being discussed. Such analysis may help highlight the dominant viewpoint and current trend. To provide this support, the system considers features such as the time-varying direction of the sentiments, the number of participants engaged in the discussion, the relationships between public sentiments and participants, and the relationships of relevant topics. The research and design questions addressed in this study focus on the following.

Q1: How to represent the sentiment characteristics of one public topic?

Q2: How to represent the time varying trend of the number of participants and their sentiment profiles?

Q3: Is there and how to represent the relationship between participants' sentiment profiles and evolution of public topics?

Q4: Can external interventions, such as from government or opinion leaders, influence participates' behavior and how can these interventions be identified and represented?

Q5: Are there any relationships between participants' interests and public sentiments and how can these relationships be identified and represented?

Q6: Can the evolution trend of public opinion be predicted from available posts and related data?

To address these questions, SentiView combines uncertainty modeling and model-driven adjustment. New sentiment mining and prediction techniques are based on text segmentation and cellular automata. Using the notion of helix with attribute astrolabe and relationship map, it supports interactive visualizations of the time-varying sentiments of participants. This paper makes the following contributions.

1) A novel model for public sentiment and its evolution for internet popular topics, using natural language processing and cellular automaton techniques.
2) A visualization method to display sentiment evolution and relationships between different participants using the helix with astrolabe and relationship map.

Manuscript received July 18, 2013; revised September 3, 2013 and October 1, 2013; accepted October 3, 2013. This work was supported in part by the Natural Science Foundation of China under Grants 61070128 and 61272199, the Natural Science Foundation for Distinguished Young Scholars of China under Grant 60925008, 973 Program of China under Grant 2010CB731402, Innovative Research Groups of the National Natural Science Foundation of China under Grant 61021004, the Innovation Program of Shanghai Municipal Education Commission under Grant 12ZZ042, the Fundamental Research Funds for the Central Universities, and Shanghai Knowledge Service Platform for Trustworthy Internet of Things under Grant ZF1213. This paper was recommended by Editor-in-Chief E. J. Bass.

C. Wang, Z. Xiao, Y. Liu, and Y. Xu are with the Software Engineering Institute, East China Normal University, Shanghai 200062, China (e-mail: cbwangcg@gmail.com; stoneshui@gmail.com; yhliu216@gmail.com; little_xu@live.cn).

A. Zhou is with the Shanghai Key Laboratory of Trustworthy Computing and Software Engineering Institute, East China Normal University, Shanghai 200062, China (e-mail: ayzhou@sei.ecnu.ediu.cn).

K. Zhang is with the School of Software Engineering, Tianjin University, Tianjin 300072, China and also with the Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: kzhang@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2013.2285047

[1] http://www.tianya.cn

[2] Statistics was taken from http://www.alexa.cn.

2168-2291 © 2013 IEEE

3) A parameterized adjustment approach to predict the time-varying evolution trend of public sentiments on popular topics.

The remainder of this paper is organized as follows: Section II summarizes related work. Section III describes our method of sentiment analysis on public topics. Section IV presents the interactive visualization for time-varying information. Section V discusses how we adjust the sentiment-driven forum data for enhanced analysis and prediction. We present experimental results and discussion in Section VI. We conclude in Section VII.

II. Related Works

There is a growing interest in visualizing sentiments from Web posts and related content. Chen *et al.* [1] presented a visual analysis system using multiple coordinated views, such as decision trees and terminology variation, to help users to understand the dynamics of conflicting opinions. Wanner *et al.* [2] described a concise visual encoding scheme to represent attributes, such as the emotional trend of each RSS news item. Both works for analyzing text contents are efficient by using word matching methods. However, they lack semantic analysis. Draper *et al.* [3] developed an interactive visualization system to allow users to visually construct queries and view the results in real time. For sentiment mining and analysis, Gregory *et al.* [4] proposed a user-directed sentiment analysis method to visualize affective document contents. Although they analyze and visualize emotion, they only use statistical methods. To demonstrate and predict the trend for an event, we suggest that rules about the evolution of public sentiments related to the participants about hot topic types should be modeled and discovered.

Collective behavior has many characteristics, such as being spontaneous, zealous, unconventional, and transient. Sentimental contagion and imitation are the main psychological mechanisms of the collective behaviors. Hoyst *et al.* [5] and Sznajd-Weron *et al.* [6] proposed two different opinion dynamics models using the aforementioned theory. For example, when discussing a debatable topic on forums, some participants' sentiments can easily be affected by others, which might result in booing or other extreme actions. In this study, we identify the changing trend of an author's sentiment from his/her posts.

Many have focused on social network visualization. Rios *et al.* [7] described how visualizations about the evolution of events on Twitter are created by presenting several case studies in recent years. Dork *et al.* [8] provided an interactive multi-faceted visual overview of large-scale ongoing conversations on Twitter, including a spiral to present participants and their activities and an image cloud to encode the popularity of event photos by size. Wu *et al.* [9] presented an interactive visual system, OpinionSeer, to analyze the collection of online hotel customer reviews by augmenting scatterplots and radial visualization. The opinion mining method can also be used in microblog sentiment mining. Baur *et al.* [10] provided an interactive visualization, LASTHISTORY, to display musical listening histories and context representing one's past. However, these methods are designed for hotel customers' feedback and music listening histories, rather than for sentiment analysis. RadViz [11], [12] was often used to map data from an n-dimensional space onto a 2-D plane, to show these features in a multidimensional space. For a complex social network, these visualization models do not focus on sentiment analysis.

Several approaches focus on the visual exploration of blogs, forum posts, and Web logs. Adnan *et al.* [13] used frequent closed patterns to model and analyze data, and create a social network. They also analyzed Web logs by integrating data mining and social network techniques [14]. Indratmo *et al.* [15] visualized Web tags and comments arranged along a time axis. Dork *et al.* [16] provided faceted visualization widgets for visual query formulation according to time, place, and tags. Ong *et al.* [17] proposed an interactive Web-based tree map, News map, to represent the relative number of articles per news item. Fisher *et al.* [18] found the evolution of topical trends in social media by using line graphs indicating term trends. The aforementioned works focus on social networks, text analysis and knowledge representation of social networks to analyze microblog and forum content without sentiment analysis.

Rose *et al.* [19] represented the change of stories by clustering keywords into themes and tracking their temporal evolution. Neviarouskaya [20] presented SentiFul to automatically generate and score a new sentiment lexicon. Lin *et al.* [21] detected sentiments and topics simultaneously from text using the JST model that is based on LDA. Zhang *et al.* [22] analyzed the sentiment of restaurant reviews in Cantonese (an important dialect in some regions of Southern China) using classification. Zhang *et al.* [23] represented a sentiment analysis method on Chinese reviews. The visualization techniques exploring the idea of live-updating views include the encoding of data changes as animations [24], and representing changes in tag frequencies [25]. The aforementioned approaches mostly provide analysis and visualization on a single sentiment aspect. Sentiment analysis alone cannot discover the law of hot topics on microblog and forum. We believe that sentiment analysis with different visualization perspectives would be more useful for users to find and understand sets of topics. Cao *et al.* [26] presented a method to show real-time information diffusion from Twitter using multiple viewing options. It traces information diffusion but ignores the relationships between different users and different hot topics, different timeframes and regions.

While there have been promising advances on visualizing the development of topics over time, research in model-driven visualization of public sentiments remains an open area.

III. Analysis of Sentiment on Public Topics

A. Sentiment Mining

Specifically, we perform sentiment modeling for each comment and sum the scores of positive and negative sentiments. This process consists of three steps described next.

First, we used spider software to analyze URLs and page tags on the forum to obtain the content and store it in a database. According to the characteristics of public topics, we collect the time, original posters, titles, content, replies, clicks, follow-up authors, replying time, and replying content. We also collect each participant's profile, such as username, gender, age, location, and signature, which can be used to classify participants and analyze their relationships.

Second, we use a HHMM-based Chinese lexical analyzer called ICTCLAS to divide these comments into several words [27]. ICTCLAS consists of three steps. The first step is atom segmentation, which divides a sentence into single words. For example, assume an original sentence is "He said it quite confidently," it becomes "##/he/said/it/quite/confidently/##" after atom segmentation. The second step is to find all possible combinations of segmented words. Finally, a word-dividing algorithm of Nth shortest path is used to find the best path, which is defined by segmentation combinations. The path refers to different selections for all possible combinations of segmented words. The binary word dividing diagram [27] is used here to couple each segmented word with its semantically most related segmented word, represented as two nodes connected by an edge.

Third, we sum the sentiment score of each segmented word to obtain the total score for the entire comment. Here, we use the HowNet sentiment analyzing dictionary [28], which includes approximately 5000 positive and 5000 negative segmented words. Each segmented word has a sentiment score. Positive sentiments are scored within [1, 0] and negative ones within [–1, 0]. For example, we can set "delight" to 0.9, "care" 0.8, "sad" –0.6 and "tragedy" –0.8. Then, we calculate the positive sentiment value of the comment by using

$$P = \sum_{i=0}^{n-1} S_i / n \tag{1}$$

where n is the number of segmented words and S_i is the score of segmented word i. If a phrase cannot be found in the HowNet segmented word dictionary, we set its score to 0. There are three types of qualifiers here: 1) enhanced: such as "very," "greatly"; 2) negative: such as "no," "not"; 3) fuzzy words: such as "possible." If such a segmented word exists, we can set a weight value for its next word. Here, the sentiment mining method is based on key words, which may be unable to find the exact sentiment within a long paragraph [29]. Since our analysis focuses on hot events involving a large number of participants, the obtained dataset is large enough and the overall sentiment prediction should be reliable.

B. Sentiment Evolution Modeling

A model that is based on cellular automaton model [30] is established to simulate the process of sentiment evolution. To analyze and predict sentiment changes, a dynamic system with discrete time and space is used. Each cell in the discrete lattice with a finite discrete state follows similar evolution rules, and is updated synchronously according to certain rules and boundary conditions. Simple interactions among a large number of cells would show the evolution of the dynamic system. It can model the interplay between adjacent participants.

First, we build a 2-D space with an $N \times N$ grid, and regard the individuals participating in the discussion of a public topic as cells with sentiment description. The developing trend can be modeled as the state evolution of these cells. There are three possible states for each cell: no comment, a supporting comment, and an opposing comment. Emotional Parameter (EP) measures the degree of sentiment tendency, including support, oppose and neutral. For a specific event, $\mathrm{EP}_i(t)$ for a cell at time t can represent the state of cell i in terms of a certain viewpoint. Assume $\mathrm{EP}_i(t) \in [0,1]$, then when $0 \leq \mathrm{EP}_i(t) < 0.5$, cell i is in the opposing state. The smaller $\mathrm{EP}_i(t)$ is, the stronger the opposition. When $0.5 < \mathrm{EP}_i(t) \leq 1$, cell i is in the supporting state. The larger $\mathrm{EP}_i(t)$ is, the stronger the support. When $\mathrm{EP}_i(t) = 0.5$, cell i is in the neutral state, neither supporting nor opposing.

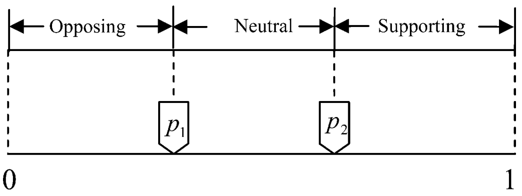


Fig. 1. Thresholds of sentiment tendency.

In general, many net surfers have the habits of only browsing social networks. Not all of them, however, will make comments to Web posts. To model this phenomenon, we introduce the concept of sentiment tendency thresholds p_1 and p_2 that can be set between [0, 1] and $p_1 < p_2$, as shown in Fig. 1. When the value of personal sentiment tendency is smaller than p_1, he/she may make an opposing comment. If the value is larger than p_2, he/she may make a supporting comment.

Let $m_i(t)$ denote the sentiment offset of cell i at the time t, defined as

$$m_i(t) = \mathrm{EC}_i(t) \times (2 \times \mathrm{EP}_i(t) - 1) \tag{2}$$

where EC is the emotional capacity to measure how independent a participant is. The higher a participant's EC is, the less he/she is influenced by his/her neighbors and external stimuli. Another interpretation is that the higher participants' EC is, the more comprehensive he or she understands the popular topics. Therefore, the person is less influenced by the others.

The change of sentiment offset precisely reflects the sentiment tendency. It can be obtained using the evolution rules between the states of neighborhoods. The factors influencing the sentiment offset in the process of the sentiment evolution include the following.

1) Fading over time: An event may gradually fade out of the public interest, resulting in a corresponding reduction of sentiment offset. Such a reduction over a period of time can be described with an exponential function as following:

$$m_i(t)' = m_i(t) \times (1 - \alpha^{m_i(t)/20}) \tag{3}$$

where α is a parameter that determines the declining speed.

2) Influence between neighborhoods: Assume that β is the parameter of neighborhood influence, and can be set on the condition of different hot topic types. Assuming n_1 participants have made comments at the time t, the influence

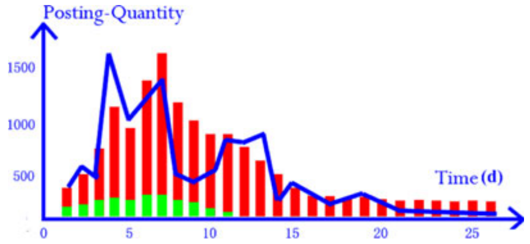


Fig. 2. Simulated number of posts (line) and real data (bars).

can be modeled by

$$m_i(t)'' = \beta \times \frac{1}{n_1} \times \sum_{i=1}^{n_1} m_i(t). \tag{4}$$

Thus, the sentiment offset of cell i at the time $t+1$ is

$$m_i(t+1) = m_i(t)' + m_i(t)''. \tag{5}$$

3) Social emergencies and external stimulus: An individual's interest in a public topic may be influenced by not only other participants, but also incidents or external stimuli. The interest can be set by modifying the boundary conditions of the cellular automata model. For example, when a positive comment appears, the number of people who tend to make a supporting comment will increase, which can be measured by reducing the value of p_2 around the grid boundary in the cellular automata model.

Fig. 2 shows the simulated results of participants in one public topic: "A debatable zero paper in China's college entrance examination," where the horizontal axis represents time and the vertical axis represents the number of posts. During simulation, we use a 100 × 100 grid, where each cell represents one participant. Initially, assume 5% cells participating in a hot topic that means 5% of the public pays attention to this topic and participate in the discussion. These participants making supporting comments account for 90%, and the individual sentiment value EP$_i(t_0)$ distributes randomly within the range of (0.5, 1]. On the other hand, the number of participants who make opposing comments is 10%, and the individual sentiment point EP$i(t_0)$ distributes randomly within the range of [0, 0.5). After about 30 iterations, the simulated data are obtained.

As shown in Fig. 2, the blue line represents the number of posts. The green bars are the number of opposing posts, while the red bars are the number of positive posts in the real data. There are up to 1700 participants in this public event. While the exact match between the simulation results and real data is not close (for example, the peak at 4 is too high and the valleys at 8 and 9 are too low), the general agreement is sufficient for our purposes.

Analyzing the real data by sentiment mining is a first step that helps to estimate the thresholds of p_1 and p_2, and initializes the proportion of participants who make supporting or opposing comments. Then, the prediction that is based on the sentiment evolution model is made. Sentiment mining is also used for our sentiment prediction and modification in Section VI.

IV. Interactive Visualization of Time-Varying Information

Public sentiments as well as the number of the participants change over time and many factors affecting the sentiments. Temporal changes could reflect the general trend of a public topic but temporal features alone are insufficient to represent the correlations among corresponding multiple factors. We, therefore, propose a multilevel visualization framework to show them simultaneously.

At the highest level, we visualize sentiment evolution along a time line. We find that a helix has an interesting structure that is suitable for this purpose. Using a helix structure, we can highlight the evolution of the number of participants, as well as their sentiments.

Having viewed the sentiment evolution over time, the user may be interested in seeing the general profile of the participants involved in the sentiment evolution. We propose attribute astrolabes to visualize multiple attributes in a single circular view, as it couples very well with a helix. We, therefore, use the two visualization tools (sentiment helix with attribute astrolabe) to interactively show the public sentiment and other characteristics over time.

To further reveal individual participants' sentiments and their relationships at the lowest level, we propose the relationship map of personal sentiment. The three visualization approaches at different levels are presented next.

A. Time-Varying Helix With Astrolabe

The tendency of public sentiments and the relevance between public sentiments and participants' attributes may be discovered by mining the collected datasets. When discussing hot topics on the Web, group sentiments are always an important indicator of the development of public events. The difficulty, however, lies in quantifying and representing sentiment indicators. We propose a sentiment helix with attribute astrolabe to visualize the public sentiment.

1) Attribute Astrolabe: To compare the attribute distribution of all participants in a given time period, attribute astrolabe maps several attributes, such as gender, onto a chart and uses the positions of points to express the attributes. The position and the color of each point indicate the attribute of one participant. The outer ring is color coded to indicate how the attribute points should be continuously, rather than discretely, distributed within the circle. As an example, we map four sets of attributes to a four-axis chart and represent them in a normalized range.

Each point represents an element of a dataset and a region can accommodate many points. Each axis represents one attribute. Too many axes would clutter the visualization space and we, therefore, recommend no more than eight axes. The position of each point can be calculated as the center of the astrolabe' gravity, which may, however, lead to the points too concentrated around the center according to the center of gravity equation. To avoid this, we amplify the variation differences among the points by squaring the distance to the center. Each point u_i has m attributes represented by the vector $\boldsymbol{S}_j(j = 1, 2, \ldots, \mathrm{m})$ and arranged to be equally spaced around the circumference of the

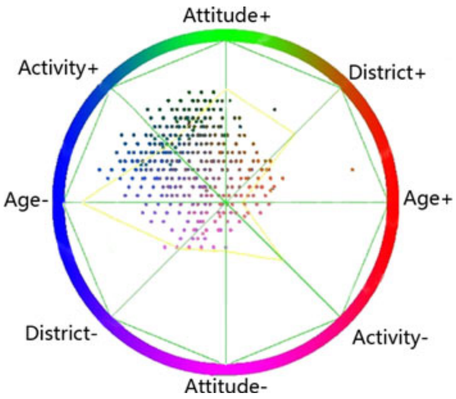


Fig. 3. Attribute astrolabe.

unit circle. The $x_{i,j}$ refers to the jth attribute value. Then, the position of u_i is calculated as follows

$$\sum_{j=1}^{m}(S_j - u_i)x_{i,j} = 0. \tag{6}$$

In RadViz [12], an m-dimensional attribute is projected to a 2-D space by (4), whose time complexity is *O(m)*. It is, therefore, *O(m × n)* to calculate n points. However, our mapping approach uses the following equation:

$$\begin{cases} x_n = \dfrac{|x_n - x_0|}{x_n - x_0} \cdot (x_n - x_0)^2 \\ y_n = \dfrac{|y_n - y_0|}{y_n - y_0} \cdot (y_n - y_0)^2 \\ z_n = \dfrac{|z_n - z_0|}{z_n - z_0} \cdot (z_n - z_0)^2 \end{cases} \tag{7}$$

where time complexity is only $O(n)$ for n points. Considering that there may be millions of datasets, the calculation efficiency is greatly enhanced. In (7), n is the number of participants for one online public topic. The coordinate value x_{ij} shows the relative distance between a point and eight sides (as indicated in Fig. 3). For each element, each value of attributes should be normalized to (0, 1) before being mapped into attribute axes.

The attribute astrolabe can show both the average trend of public sentiment, and the individual attributes of participants. In addition, multiple astrolabes can be used to compare the variations in different time periods.

As shown in Fig. 3, one colored point in the attribute astrolabe represents the attribute of the participant. Different colors are for different attributes. The relative distances between this point and eight sides show the participant's characteristics. For example, if the point is near the side of Age+ toward the left, this is an older participant. If the point is near the side of District+, the average income of the district where this participant lives is higher than those on the other side.

2) Sentiment Helix: To illustrate the time distribution and evolution of all participants, a helix uses the rotation angle to represent the average sentiment tendency, as shown in Fig. 4. The ascent of the helix shows the overall tendency of public sentiment over time. Because of the tubular shape of the helix, its diameter can be used to show the number of participants. By locating attribute astrolabes in different time periods next to the spiral axes, the evolution trend can be tracked and analyzed.

Fig. 4 visualizes the event of "3Q WAR," a conflict between two well-known Chinese IT companies, Tencent (well known for QQ IM) and Qihoo (well known for 360 antivirus). We collected over 20 000 posts from the Tianya forum and the ages of participants, online activity, sentiment tendency, and local features from statistical and semantic analyses. We divided the information into eight parts and normalized them.

Fig. 4(a) demonstrates that the opinions for "3Q WAR" became polarized after six popular IT companies, i.e., Kingsoft, Baidu, Sohu, Keniu, Mathon, and Rising, joined the debate. They are divided in two groups that are, respectively, composed of the supporters of Tencent and Qihoo. Then, the public attitude shifted from supporting Qihoo to supporting Tencent. The helix goes down before the fifth day and up afterwards. In addition, we can observe the users' participation level by the changing size of the helix diameter. After about a week, the smaller helix indicates that the number of participants in the later-stage dropped. They may no longer be interested in this topic, and thus, the public attention would change.

We also collected about 200 million items of data on the same topic from the Sina microblog. Based on the characteristics of the blogs, we extracted four sets of the attributes: number of forward blogs (i.e., new blogs following other blogs), number of commented blogs (i.e., blogs commenting on others blogs), total number of forward for the original blog, and total number of comments on the original blog. The two latter are defined as below.

If a is the forward blog for b, annotated as $b \rightarrow a$, then

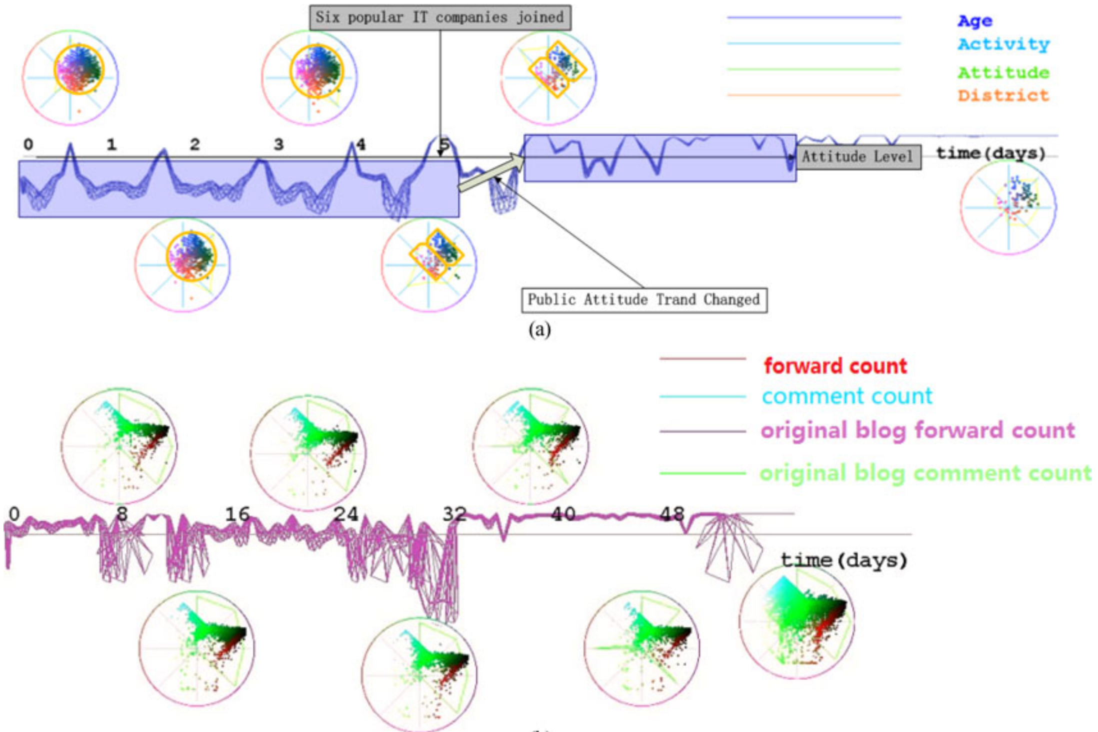
$$\mathrm{FC} = |\{a|\forall a\colon\ b \rightarrow a\}| \tag{8}$$

is the total number of forwards (or forward count) from b. If $b = b_0$ is the original blog, FC$_0$ is the total number of forwards for the original blog. If a is the comment blog on b, annotated as $b \Rightarrow a$, then

$$\mathrm{CC} = |\{a|\forall a\colon\ b \Rightarrow a\}| \tag{9}$$

is the total number of commented blogs (or comment count) on a. If $b = b_0$ is the original blog, CC$_0$ is the total number of commented blogs on the original blog.

The results in Fig. 4(b) show that there is more attention in forwarded blogs and comments. The more points appearing blue and red, the more participants post new blogs and comments. The participants in these blogs with social influence or popularity are always VIP users (who are marked and verified by Sina). At the beginning, there are fewer participants but more VIP participants than in the middle period as inferred by contrasting Fig. 4(b) and raw data. After the beginning, the blogs have fewer forwards and comments but an increasing number of the original blog forward count (to the right) and original blog comment count (to the lower part). It indicates that many ordinary participants appeared when the public attention was attracted to that event. A sharp decrease emerges on general blogs after the decrease of frequently forwarded and commented blogs. By studying the evolution, we can find that in "3Q WAR," the



(a)

(b)

Fig. 4. Sentiment helix with attribute astrolabe. (a) Forum information. (b) Sina microblog information.

public's attention was triggered by VIP users and changed as soon as the VIP users changed their attitudes.

B. Relationship Map

In public opinions, personal sentiments could influence each other, and people may be characterized by different sentiment features. Exploring the hidden relationships between different participants helps the analysis.

The relationship map highlights the participants who have a common interest in the forum and facilitates the analysis of sentiments expressed on hot topics. We divide sentiments into three categories: positive, negative, and neutral.

First, we preprocess the data input from a Chinese lexical analyzer. Any incomplete data, such as those missing attributes (e.g., age, comments), are removed. We then break the comment sentences into words, and count the positive and negative words to obtain the scores of positive or negative comments. We propose a visualization method that consists of topic ellipses, participants drawn as points and links as relationships.

1) Topic Ellipse: Each ellipse includes replies for one topic. The area of the ellipse corresponds to the total number of comments on the topic, and its shape is determined by the ratio of the total number of comments to that of distinct participants involved (as a participant can make more than one comment). The ellipse reflects the coverage of a topic among different participants. In the ellipse in Fig. 5, there are two small ellipses, the left one contains positive comments and the right one contains negative comments. Other areas in the big ellipse contain neutral comments. Similar to the big ellipse, the size of the small ellipse corresponds to the number of comments with a certain type of sentiment and the color represents the average number of words in the comments.

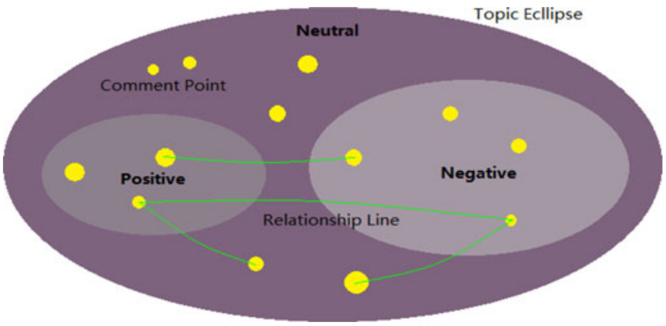


Fig. 5. Relationship map.

2) Comment Point: The size of a point represents a poster's total number of words with a sentiment (positive, negative, or neutral). The position of the point indicates the post's sentiment inclination. The points in the large eclipse contain posters, who have expressed neutral comments.

3) Relationship Line: A green line is used to link the same poster in positive and negative ellipses. The lines start with red and end with blue [as shown in Fig. 6(a)] to indicate the relationship that the authors with red color pay much attention to the information of blue points. Line bundling is used for an esthetics reason. We perform the Bezier curve algorithm to draw the lines.

4) Interactions: When first loaded, all comments from the same posters are connected in light green lines. If the user clicks on a comment, the point for that comment will be highlighted

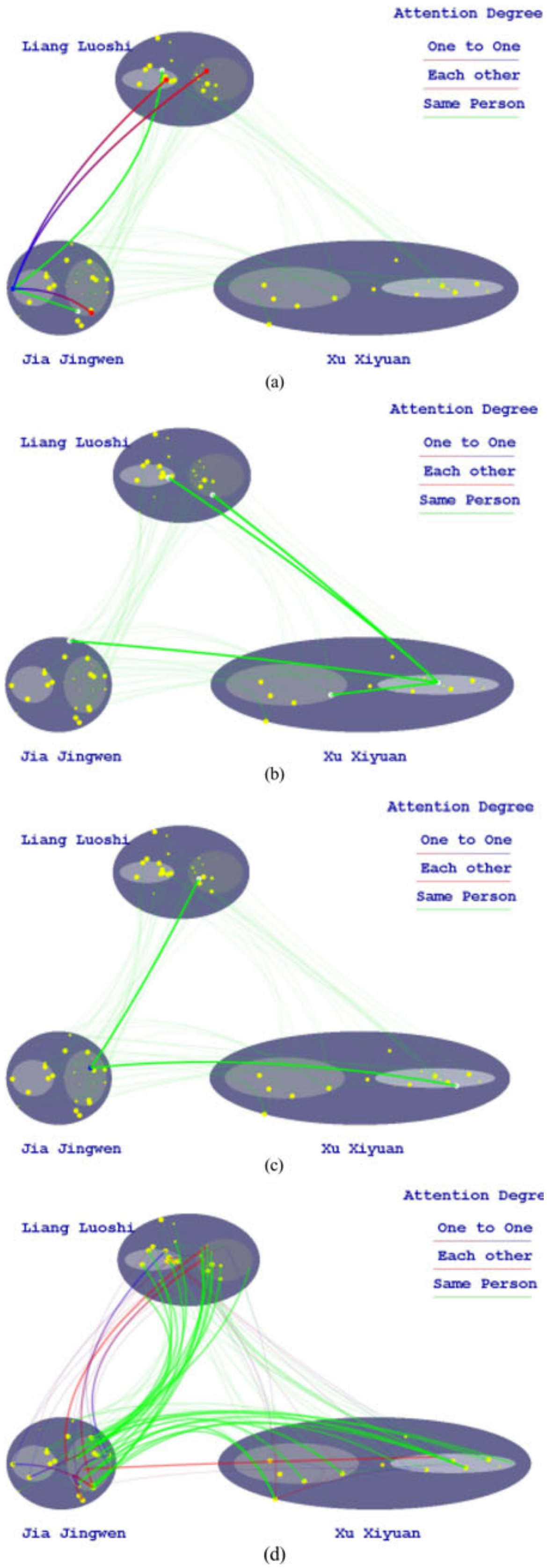


Fig. 6. Sentiment relationship maps showing different types of participants.

in green, and all other comments related to it will also be highlighted (such as those by the same author, those followed by the current author, and those whose authors are fans (Somebody's followers are called their fans in Sina blog) of the current author).

Fig. 6 shows a relationship map for the hot topic: "Many actresses dating the rich and powerful," with varied types of posters. Each comment point may only be connected to the ones made by the same participant, or only related with those who share interests. By analyzing different participants with their follow-up relationships, we can reach a conclusion whether people with similar interests may have similar personal sentiments. At the same time, we can obtain interesting attributes about the participants involved in different posts on the same theme. The viewpoints of participants are relatively balanced [see Fig. 6(a)]. The number of positive comments almost equals the number of negative comments in the three events. Fig. 6(b) shows one type of participants, with a sentiment tendency toward each post, but that vary among different posts, yet focused on similar themes. The bold line shows the same participants joined in the three topics, but the comments of the participants have different sentiments, visualized around the corresponding three topic ellipses. On the contrary, in Fig. 6(c), another type of participants holds similar views by posting negative comments on the three events.

We can also show relationships among special groups without clicking on any points. In Fig. 6(d), using statistical results, 10% of shortest but negative comments on the topic of "Jia Jingwen" at the bottom left are highlighted. There is a negative tendency as more negative comments were posted than positive comments.

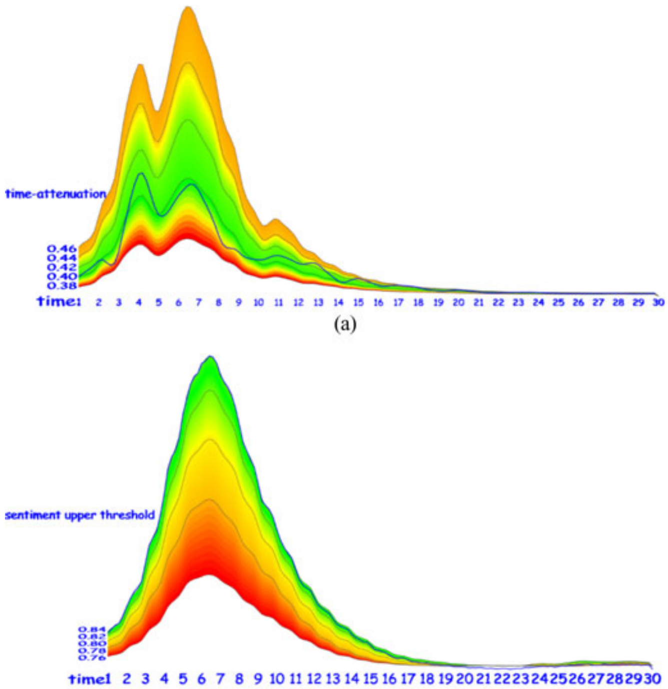
V. Adjustment of Sentiment-Driven Forum Data

The aforementioned visualization and analysis are all based on the real online data. Sometimes such real data are incomplete. Being able to predict the evolution of public topics that are based on the existing data is particularly useful. Using the sentiment prediction model in Section IV, we can further simulate these characteristics and sentiment changes on public topics.

This section's analysis is based on the data that are obtained on three hot topics: "A debatable zero score paper in China's college entrance examination," "3Q War," and "Many actresses dating the rich and powerful."

Simulated results might be inaccurate, as they may be parameter sensitive. We, therefore, adjust the parameters and identify their optimal ranges by comparing and matching the simulation data with the real data.

We first define two variables: 1) deviation, measuring the difference between the simulation value and the real value at every moment, and 2) reliability representing the accuracy of the simulation results, measured by the average deviation. We can enhance the reliability by adding the actual deviation values with the corresponding weight values. When the number of posts is small, for example, at the beginning and end phases in Fig. 2, small weight values are set for the actual deviations to enhance the reliability.



(a)

(b)

Fig. 7. Visualization of the adjustment on the simulation data against real data. (a) Verifying the time-varying parameter (zero score event). (b) Adjustment the sentiment threshold p_2 (3Q war event).

There are many time-varying visualization methods [31]. Here, we use a theme—river-like diagram [32] to represent different reliability values. As Fig. 7 shows, the horizontal axis represents time measured in days and the vertical axis is used to encode the number of posts corresponding to the parameter being measured in the prediction model. Fig. 7 shows a time-varying parameter and a parameter for the sentiment threshold p_2, as discussed in Section IV. We obtained different simulation results for the posting quantity using different parameter values in the prediction model. The vertical width of an individual stream is determined by the posting quantity during the specific time interval corresponding to the parameter value on the left side. All streams in different colors are stacked vertically to represent the reliability values under different simulation parameters. The higher the reliability is, the greener the color. Similarly, lower reliability is represented by reddish colors.

The blue line represents the posting quantity over time for the real data, which is set near the layer corresponding to the optimal parameter. We collect the real data from Tianya Forum. The posting quantities of some hot topics in one month are counted. Taking Fig. 7(a) as an example, the color encodes the reliability value, indicating that the posting quantity simulation at the parameter 0.40 has the highest reliability.

If the real data are incomplete or if we wish to predict the evolution of public topics in the near future, we can use our sentiment prediction model with the optimal parameters to simulate the incomplete or future data. This is an advantage of adjustment-based visualization using real data.

Fig. 8 shows an example of an incomplete real dataset, missing data after the seventh day. The reliability value in the posting quantity simulation is the highest with the parameter 0.36. The results can be used to predict the future trend by compensating the missing data with the simulated ones.

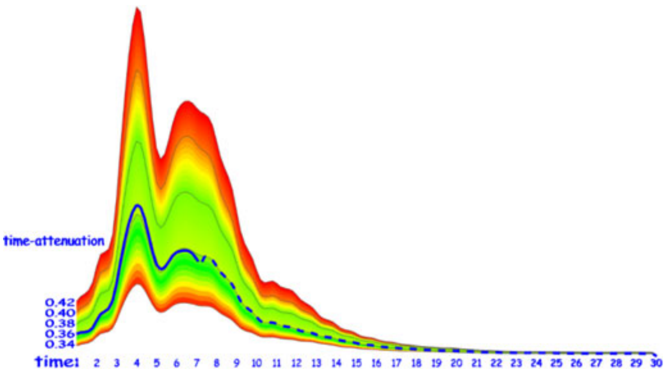


Fig. 8. Prediction based on real data and simulated results (Actresses dating event).

VI. Experimental Results

The visualization in this paper was developed using Microsoft Visual C++ 2008, OpenGL library, and QT library. The experiments were run on a PC with an Intel Pentium Dural Core processor, DDR2 2G RAM and Nvidia 7900 GS display card.

The performance of our system as compared with related ones appears in Table I. This comparison is made with the same development environment for all approaches. Our visualization is more geared toward representing time-varying trends with appropriate adjustment. It is also faster with respect to rendering speed. We could not obtain the speed data on StoryVis [19] and VisGets [16]. Since StoryVis' multidimensional visualization is implicit (although it is unable to verify the sentiment trend). VisGets has multidimensional visualization without time-varying capability.

The three aspects of our visualization approach complement each other. First, the sentiment helix with astrolabe focuses on the time-varying evolution trend and attributes comparison between individuals. Second, the sentiment relationship map can visualize the interest and sentiment correlation among different participants. Third, the enhanced river-like diagram can visualize and verify simulated results based on real data, useful for effectively managing and predicting the sentiments on Internet topics.

Using the system, we collected data for several hot topics such as "3Q war event" and "zero score event" in various forums and analyzed them. In order to gain feedback on the usefulness and limitations of SentiView, more than 1000 people were invited to use SentiView, and then, to complete a questionnaire sent via email. 300 people responded to our questionnaire including college students, scholars, government officers, police officers, employees of advertising agencies and unknowns. Their ages ranged between 16∼53. All were of Chinese origin from over ten provinces of China. We divide the participants broadly into three categories: those who actively participated in public online discussions, those who watched the discussions, and govern-

TABLE I
PERFORMANCE COMPARISON

	Time-varying	Multi-Dimension	Adjustment	Render Speed (frames per second)
Radviz [11][12]	No	Explicit	No	Slow (<10fps)
StoryVis [19]	Yes	Implicit	No	Not available
VisGets [16]	No	Explicit	No	Not available
SentiView	Yes	Explicit	Yes	Fast (15~24fps)

TABLE II
QUESTIONNAIRE AND FEEDBACK FROM THE SUBJECTS

1. What's your profession?	IT	Financial	Manager	Other
	134	56	15	95
2. Can you identify the life cycle of public topic and evolution from the sentiment helix?	Yes. I can easily figure them out.			I can only find one (or none) of them.
	292			8
3. Can you find different attributes of participates from the time-varying astrolabe?	I can see the number of VIP users decreased with the decrease of general users.			I can only see the number of all users decreasing.
	233			67
4. Can you see the tendency of topics' sentiment and participants from the sentiment relationship map?	The points of participants and the ellipses can obviously show them.			It is difficult for me to understand all of it.
	280			20
5. Based on the simulation results, can you easily forecast the development of topics?	Yes. I can.			It is difficult for me to forecast from the simulation results. Or I have no idea.
	256			44
6. Is this system useful?	Very useful			A little use/No use
	289			11

ment policy makers. The questionnaire and collected answers are reported in Table II. We summarize the feedback next.

A. People are Actively Participating in the Public Topics

The users believed that SentiView could clearly show the trend of an entire event and the types of participants. For example, for the event of "Feng sister," at the beginning a great number of new participants paid attention to the topic for a short period of time. Among them, the youths were in the majority. Using SentiView, they could easily find which topics are worthy of attention. This helps to determine if it is beneficial to participate in the forum in the future.

B. Network Monitoring Personnel Identified Patterns

Network monitoring analysts felt that this system could assist in monitoring for patterns in the Internet hot topics and in considering different external factors that impact the development of public sentiment. For example, a general manager commented, "When the conflicts of QQ and 360 just occurred, many VIP users participated; later a large number of general users followed and participated in the argument. Then, we began to mediate Tencent and Qihoo, when most VIP users shifted their focus away. Then, even after more general users joined, the topic would still become obsolete. To my surprise, the system showed a pattern so similar to what exactly happened. By clicking on points in the relationship map, one can check his/her own position on specific topics as well as find which topics other participants are interested in as well as their opinions on the topics."

C. Government Policy Makers Identified Shifts in Public Sentiment

Policy makers used SentiView to discover the evolution of hot topics in the forums and the current trend of public viewpoints. By identifying public views and attitudes, policy makers could steer events in the direction of their expectations. Particularly, noteworthy was the prediction that is based on real data.

Furthermore, one of the potential users provided recommendations about components of the system that should be integrated into one program.

VII. DISCUSSION

This paper has introduced a new visualization system for analyzing, visualizing and verifying the sentiments of Web users on public topics. A text-based sentiment mining method and a model-driven prediction approach have been used to analyze the public sentiments on hot topics (Q1). Considering the characters and interests among different participants in some popular topics, we have proposed two new visualization concepts, the helix combined with astrolabe (Q2, Q3, and Q4) and relationship map (Q5) to visualize changes of multiple attributes and complex relations among the attributes, such as numbers, location distribution, ages, and sentiment, of the participants. Furthermore, using an evolution model for simulation, we can compare time-varying parameters in the simulation against real data on sentiment-driven forum information (Q4 and Q6). Since our analysis approaches can be tailored to meet different requirements, SentiView can be used to analyze and visualize mass Web information effectively in many applications.

SentiView builds upon and extend several ideas from state-of-the-art techniques to enable advanced visual analysis of public sentiments on popular topics on the Internet. Attribute astrolabe is adopted from RadViz [11], [12] that maps data from an n-dimensional space onto a 2-D plane, to show these attributes in a multidimensional space. However, compared with the mapping method that is based on physics in RadViz [11], [12], our mapping method in the attribute astrolabe is based on the geometry, and thus, more efficient. Sentiment helix extends line graphs [18] that use line heights to represent temporally changing data. The ascent of a helix shows the overall tendency of public sentiment over time, while its width represents the number of participants. The relationship map highlights the participants who have a common interest in the forum and facilitates the analysis of sentiments expressed on hot topics, similar to network graphs. The theme—river-like diagram is inspired from the ThemeRiver [32] to represent the comparisons of simulated results under different simulation parameters. Three system components designed in SentiView are showing rich information from different aspects at once and provide flexibility for varying tasks.

Acknowledgment

The authors would like to thank H. Ma, B. Zhao, and Q. Zhang for providing the data on Sina micro-blog and spider software, the anonymous reviewers, and the Editor-in-Chief for their insightful comments that have helped us to improve the presentation of the paper.

References

[1] C. Chen, F. Ibekwe-SanJuan, and E. SanJuan, "Visual analysis of conflicting opinions," in *Proc. IEEE Comput. Soc. Symp. Visual Analytics*, Chicago, IL, USA, 2006, pp. 59–66.

[2] F. Wanner, C. Rohrdantz, F. Mansmann, D. Oelke, and D. Keim, "Visual sentiment analysis of RSS news feeds featuring the US presidential election in 2008," presented at the Workshop on Visual Interfaces to the Social and the Semantic Web, Sanibel Island, FL, USA, 2008.

[3] G. Draper and R. Riesenfeld, "Who votes for what? A visual query language for opinion data," *IEEE Transa. Vis. Comput. Graphics*, vol. 14, no. 6, pp. 1197–1204, Nov./Dec. 2008.

[4] M. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner, "User-directed sentiment analysis: Visualizing the affective content of documents," in *Proc. Workshop Sentiment Subjectivity Text*, 2006, pp. 23–30.

[5] J. Hoyst, K. Kaceperski, and F. Schweitzer, *Annual Reviews of Computational Physics IX*. Singapore: World Scientific, 2001.

[6] K. Sznajd-Weron, "Sznajd model and its applications," *Acta Phys. Polonica B*, vol. 36, p. 2537, 2005.

[7] M. Rios and J. Lin, "Distilling massive amounts of data into simple visualizations: Twitter case studies," in *Proc. Int. AAAI Conf. Weblogs Social Media*, 2012, pp. 22–25.

[8] M. Dörk, D. Gruen, C. Williamson, and S. Carpendale, "A visual backchannel for large-scale events," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1129–38, Dec. 2010.

[9] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu, "OpinionSeer: Interactive visualization of hotel customer feedback," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1108–1118, Nov./Dec. 2010.

[10] D. Baur, F. Seiffert, M. Sedlmair, and S. Boring, "The streams of our lives: Visualizing listening histories in context," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1119–1128, Nov./Dec. 2010.

[11] L. Nováková and O. Štepanková, "RadViz and identification of clusters in multidimensional data," in *Proc. 13th Int. Conf. Inf. Vis.*, 2009, pp. 104–109.

[12] J. Stasko, G. Grinstein, and K. A. Marx, "Vectorized RadViz and its application to multiple cluster datasets," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 6, pp. 1444–1427, Nov./Dec. 2008.

[13] M. Adnan, R. Alhajj, and J. Rokne, "Identifying social communities by frequent pattern mining," in *Proc. 13th Int. Conf. Inf. Vis.*, 2009, pp. 413–418.

[14] M. Adnan, M. Nagi, K. Kianmehr, M. Ridley, R. Alhajj, and J. Rokne, "Promoting where, when and what? An analysis of web logs by integrating data mining and social network techniques to guide eCommerce business promotions," *J. Social Netw. Anal. Mining*, vol. 1, no. 3, pp. 173–185, 2010.

[15] J. Vassileva and C. Gutwin, "Exploring blog archives with interactive visualization," in *Proc. Conf. Adv. Vis. Interfaces*, 2008, pp. 39–46.

[16] M. Dork, S. Carpendale, C. Collins, and C. Williamson, "VisGets: Coordinated visualizations for Web-based information exploration and discovery," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 6, pp. 1205–1212, Nov./Dec. 2008.

[17] T. Ong, H. Chen, W. Sung, and B. Zhu, "Newsmap: A knowledge map for online news," *Decision Support Syst.*, vol. 39, no. 4, pp. 583–597, 2005.

[18] D. Fisher, A. Hoff, G. Robertson, and M. Hurst, "Narratives: A visualization to track narrative events as they develop," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, 2008, pp. 115–122.

[19] S. Rose, S. Butner, W. Cowley, M. Gregory, and J. Walker, "Describing story evolution from dynamic information streams," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, 2009.

[20] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "SentiFul: A lexicon for sentiment analysis," *IEEE Trans. Affective Comput.*, vol. 2, no. 1, pp. 22–36, 2011.

[21] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly-supervised joint sentiment-topic detection from text," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1134–1145, Jun. 2012.

[22] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of Internet restaurant reviews written in cantonese," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7674–7682, 2011.

[23] W. Zhang, H. Xu, and W. Wan, "Weakness finder: Find product weakness from chinese reviews by using aspects based sentiment analysis," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 10283–10291, 2012.

[24] E. Hetzler, V. Crow, D. Payne, and A. Turner, "Turning the bucket of text into a pipe," in *Proc. IEEE Comput. Soc. Symp. Inf. Vis.*, 2005, pp. 89–94.

[25] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins, "Visualizing tags over time," *ACM Trans. Web*, vol. 1, no. 2, pp. 7-es (Art. 7), 2007.

[26] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2649–2658, Dec. 2012.

[27] H. Zhang, H. Yu, and D. Xiong, "HHMM-based chinese lexical analyzer ICTCLAS," in *Proc. Second SIGHAN Workshop Chin. Language Proc.*, 2003, pp. 184–187.

[28] Z. Dong and Q. Dong, *HowNet And the Computation of Meaning. River Edge*. River Edge, NJ, USA: World Scientific, 2006.

[29] F. Wanner, C. Rohrdantz, F. Mansmann, D. Oelke, and D. A. Keim, "Visual sentiment analysis of RSS news feeds featuring the us presidential election in 2008," presented at the Workshop Vis. InterfacesSemantic Web, Sanibel Island, FL, USA, 2009.

[30] A. Nowak and M. Lewenstein, *Modeling Social Change with Cellular Automata*, R. Hegselmann, U. Mueller, and K. G. Troitzsch, Eds. Boston, MA, USA: Kluwer, 1996, pp. 249–285.

[31] W. Aigner, S. Miksch, H. Schumann, and C. Tominski, "Visualization of time-oriented data," in *Human-Computer Interaction*, 1st ed. London, U.K.: Springer, 2011.

[32] S. Havre, B. Hetzler, and L. Nowell, "ThemeRiver: visualizing theme changes over time," in *Proc. IEEE Comput. Soc. Symp. Inf. Vis.*, Washington, DC, USA, 2000, pp. 115–123.

Changbo Wang received the Ph.D. degree from the State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China, in 2006.

He is currently a Professor with the Software Engineering Institute, East China Normal University, Shanghai, China. He was a Visiting Scholar with the State University of New York, Stony Brook, NY, USA, from 2009 to 2010. His research interests include computer graphics, information visualization, and virtual reality.

Zhao Xiao received the B.E. degree in software engineering from the School of Software, Nanchang University, Nanchang, China, in 2009 and the M.E. degree from Software Engineering Institute, East China Normal University, Shanghai, China, in 2012. He is currently working toward the Ph.D. degree with the Software Engineering Institute, Tianjin University, Tianjing, China.

His current research interests include computer graphics and information visualization.

Yuhua Liu received the B.E. degree in software engineering at the Software Engineering Institute, East China Normal University, Shanghai, China, in 2011, where he is currently working toward the Ph.D. degree.

His current research interests are information visualization and computer graphics.

Yanru Xu received the B.E. degree in software engineering from the Software Engineering Institute, East China Normal University, Shanghai, China, in 2009, where she is currently working toward the M.S. degree.

Her current research interests include information visualization.

Aoying Zhou received the Ph.D. degree from Fudan University, Shanghai, China, in 1993.

He is currently a Professor in computer science with the East China Normal University, Shanghai, where he is also chairing the Institute of Massive Computing. He is now serving as the Vice Director of ACM SIGMOD China and Database Technology Committee of China Computer Federation. His research interests include massive data management, Web data management and Web mining, Web services, data streams, and P2P computing systems.

Dr. Zhou is a Member of the editorial boards of the *VLDB Journal* and the *Journal of Computer Science and Technology (JCST).*

Kang Zhang received the B.Eng. degree in computer engineering from the University of Electronic Science and Technology of China, Chengdu, China, in February 1982, the Ph.D. degree from the University of Brighton, U.K., in December 1990, and the Executive MBA degree from the University of Texas at Dallas, USA, in May 2011.

He is currently a Professor and the Director of the Visual Computing Lab, Department of Computer Science, University of Texas at Dallas, Richardson, TX, USA, and an Adjunct Professor with the School of Software Engineering, Tianjin University, Tianjin, China. His research interests include visual languages, aesthetic computing, information visualization, and software engineering.

He is on the editorial boards of the *Journal of Visual Languages and Computing, Journal of Big Data* and the *International Journal of Software Engineering and Knowledge Engineering*.