Journal of Visual Languages and Computing 28 (2015) 122–133

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Analyzing internet topics by visualizing microblog retweeting ☆

Changbo Wang [a,c,*], Yuhua Liu [a], Zhao Xiao [a], Aoying Zhou [a,c], Kang Zhang [b]

[a] Software Engineering Institute, East China Normal University, Shanghai 200062, China
[b] Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080, USA
[c] Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China

ARTICLE INFO

Article history:
Received 16 March 2014
Received in revised form
15 July 2014
Accepted 19 November 2014
Available online 15 December 2014

Keywords:
Microblog retweeting
Internet topics
Visualization
Analyzing

ABSTRACT

Microblog is a large-scale information sharing platform where retweeting plays an important role in information diffusion. Analyzing retweeting evolutions can help reasoning about the trend of public opinions. Information visualization techniques are used to demonstrate the retweeting behavior in order to understand how Internet topics diffuse on Microblogs. First, a graph clustering method is used to analyze the retweeting relationships among people of different occupations. Then a new algorithm based on electric field is proposed to visualize the layout of the relationship links. A prediction method based on three diffusion models is presented to predict the number of retweets over time. Finally, three real world case studies show the validity of our methods.

1. Introduction

Microblogging provides a new and convenient mechanism for people to share information within their virtual communities anywhere at anytime using desktop and mobile devices. Its representatives of Web 2.0, such as Twitter, Facebook, all have attracted a large number of users. More and more communications are taking place through interactive user behaviors on microblogs. Those behaviors include tweets publishing, browsing, retweeting and replying. Among them retweeting is the key mechanism for information diffusion on microblog, such as the behavior that users retweet the tweets of their friends.

Retweeting connections have been intensively studied recently on network properties and information diffusion. Yet none of the previous approaches offers multiple-view visualization for users to easily explore the trend and retweeting patterns. Existing approaches also cannot predict the number of future retweets based on the past retweeting data. To explore the structures and attributes of retweeting networks, we provide an evolving, interactive visualization interface known as MV^2 (MultiView Microblog Visualization), which integrates three visualizations (see Fig. 1), i.e. Virtual Map, Clustered Graph and Retweet Streams. They work together to describe the retweeting evolutions of an Internet topic. The Virtual Map shows the dissemination of information among different cities. The Clustered Graph shows retweeting connections among users' occupations. They are accompanied by Retweet Streams, which present the predicted and actual numbers of retweets over time. MV^2 also supports temporal exploration through the time slider at the bottom. The timeline control allows users to slide back and forth to explore data within different time in the history. Thus MV^2 offers a multi-angle visualization in presenting how a hot public opinion diffuses.

The rest of this paper is organized as follows. Section 2 provides a brief overview of prior related work. Section 3

☆ This paper has been recommended for acceptance by S.-K. Chang.
* Corresponding author. Tel.: +86 21 62224122.
E-mail address: cbwangcg@gmail.com (C. Wang).

http://dx.doi.org/10.1016/j.jvlc.2014.11.007
1045-926X/

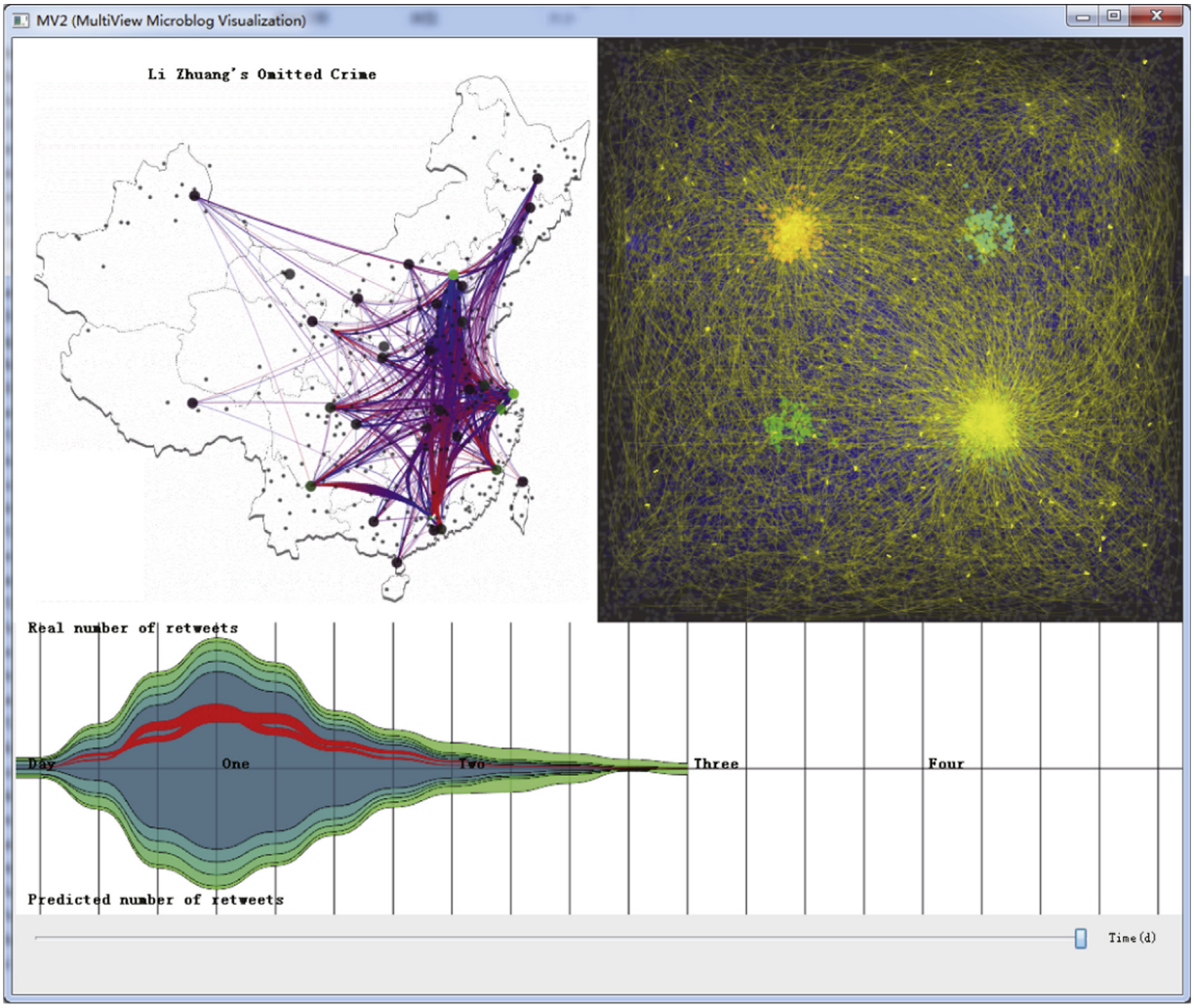

Fig. 1. Virtual Map (top left) displaying retweeting connections among major cities, Clustered Graph (top right) indicating retweeting connections among occupations, and Retweet Streams (bottom) representing the predicted and actual numbers of retweets over time.

analyses the theoretical basis of MV^2 which involves graph clustering based on attribute similarities, an electric field model for visualizing connection links after clustering, and the prediction model for retweeting numbers. Then the design of the three visualization approaches is described in Section 4. Section 5 discusses the empirical results while Section 6 concludes the paper and mentions the future works.

2. Related work

Along with the development of Microblog, most studies focus on the analysis of retweeting behavior. Suh et al. [1] examined the features that might affect the retweetability of tweets. They found that retweeting was associated with various social motivations such as entertaining a specific audience, commenting on someone's tweet, publicly agreeing with someone. Xie et al. [2] modeled the retweeting process by exploring the product of motion, identification and interests. Yang et al. [3] proposed a factor graph model to predict users' retweeting behaviors by analyzing users' regularity of retweeting, importance of tweet contents and users' interest. However, they all explore the mechanism of retweeting behavior from the view of data mining, which do not provide an intuitive overview. Our proposed approach uses information visualization to analyze large-scale retweeting behaviors related to social topics on microblogs.

In the area of graph visualization, force directed algorithms [4–7] are popular in generating graph layouts. They define a system of forces acting on the vertices and edges and finally find a minimum energy state by solving differential equations or simulating the evolution of the system. Graphs drawn with these algorithms tend to be esthetically pleasing, exhibit symmetries. They, however, only use the information contained in the graph structure, rather than the semantics of nodes and edges. In order to analyze the retweeting connections within users' occupations, we improve one of the force-directed algorithms to compute the graph layout based on vertices' attributes.

Selassie et al. [8] presented a divided edge bundling method to reduce the clutter of lines and improve the readability of node-link lines considering the direction and weight of edges. By considering retweeting connections among different users, we propose a new algorithm based on electric field to compute the layout of links that reduces the overall node collisions and edge crossings.

Most of the current social networks are displayed in a static form, without offering an overview of the evolving process. Moody et al. [9] studied a type of visualization for temporal social networks: static flip books (a combination of fixed node layout and dynamic social relationships), where the node positions remain constant but the edges accumulate over time [10]. Bender-deMoll and McFarland [11] further studied another type of dynamic movies where nodes move as a function of changes in relations [12]. By combining the above two types of visualization modes to present retweeting networks in space and time, the user could get a global sense of how a hot topic spreads. Byron and Wattenberg [13] applied and adapted the stacked graphs to entertainment datasets and suggested new methods for ordering and coloring the stacked layers. In this paper we introduce a stacked graph technique that contrasts the predicted and actual numbers of retweets temporally, topically, and dynamically.

Several approaches have been proposed for visual exploration of Twitters and microblogs by providing aggregated information. Dork et al. [14] designed a Visual Backchannel for large-scale events to provide an evolving, interactive, and multi-faceted visual overview of large-scale ongoing conversations on Twitter. Diakopoulos et al. [15] presented a visual analytic tool, Vox Civitas, for helping journalists and media professionals extract news from large-scale aggregations of social media contents around broadcast events. Hao et al. [16] designed two visual analytics tools, Pixel Sentiment Calendar and Pixel Sentiment Geo Map, to show Twitters' distribution and patterns, and to identify influential opinions. Marcus et al. [19] developed TwitInfo to automatically detect and display peaks of high tweet activity. Previous research has also used clustering to reduce data complexity and to facilitate analysis. Gansner et al. [20] described a text stream visualization method that initially groups tweets by “countries” and then generates a dynamic map. ThemeCrowds [21] displayed topic trends on Twitter over time using multi-scale tag clouds. Twitter users are clustered hierarchically and then visualized based on the topics they discuss. Recently, several visualizations have been designed to show the spread of information on social media. Google+ Ripples [22] monitored and depicted how a single post spreads over the user network and its communities. Cao et al. [23] developed Whisper to visualize the spatio-temporal process of information diffusion on Twitter. In contrast to the above approaches, MV^2 not only visualizes Twitters' distribution and social network patterns but also explores the retweeting connections among cities and communities based on these users' occupations. More importantly, we could predict and visualize the development trend on retweet numbers.

3. Analysis of microblog retweeting

To explore the potential patterns of retweeting connections among users' occupations, the analyzing process would include a graph clustering on vertices' attributes, and the layout of the connection graph. By changing the attributes, the similar clustering approach could be used to explore retweeting patterns based on other user profiles, such as age. The clustering and animation model in the following two subsections are therefore generic.

3.1. Clustering with attribute similarities

Graph clustering techniques are very useful for detecting densely connected groups in a large graph. To analyze large-scale retweeting behaviors about some social topics on Microblog, it is necessary for us to divide vertices of a retweeting network into different clusters based on the profiles of microblog users. However these standard graph clustering techniques mainly focus on the topological structure but largely ignore vertices' attributes. Here we need clustering according to the characteristics of the vertices to explore potential patterns of retweeting connections among users' occupations. Besides, to keep the inner structure of each cluster decided by the relationships among nodes, the FR-algorithm [7] is extended to obtain graph clustering by introducing a new force based on the attribute similarities of the nodes. Thus our method can obtain several clusters considering both the nodes' attributes and the topological structure to better suit the social networks. Each of vertices is randomly placed and will be grouped into clusters whose centers are selected in advance. Each center of the cluster represents one attribute and seeks to group together the vertices with the same attribute. There are attractive forces between adjacent connected vertices and repulsive forces between all pairs of unconnected vertices close to each other.

For example, as shown in Fig. 2, vertex P in the lower left corner is one of the centers of clusters. Meanwhile, different colors in Fig. 2 represent different attributes. Then we calculate the force on vertex V_0. A repulsive area is drawn as a circle whose center is V_0, and radius R is defined as $R=\sqrt{S/N}$. Here S represents the size of the area occupied by the graph and N is the number of vertices. We take R as the optimal distance between vertices. Only if the

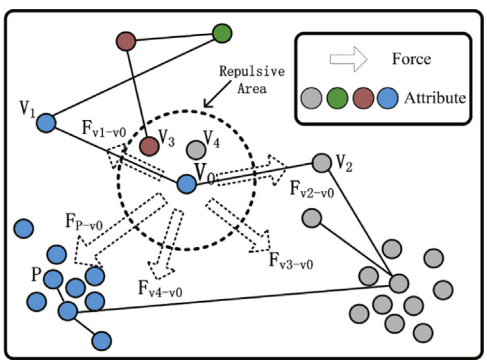

Fig. 2. Calulating repulsive and attractive forces.

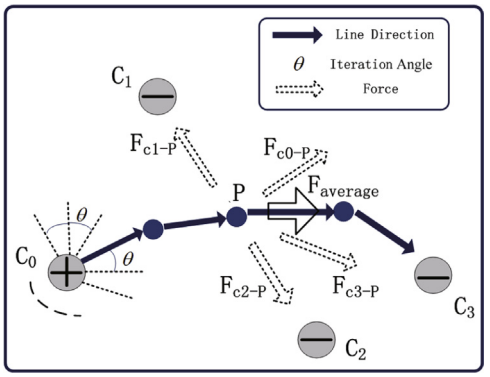

Fig. 3. Calculating the direction of connection lines using Coulomb's law. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

distance between two vertices is less than or equal to R, they repel each other. So the two vertices V_3 and V_4 in the repulsive area both exert a repulsive force against V_0. There is also an attraction between these connected vertices, for example, vertices V_1 and V_2 connected to V_0 both attract V_0. The attractive and repulsive forces can be defined as $f_a(d)=d^2/R$, $f_r(d)=R^2/d$ respectively in terms of the distance d between two vertices. Then the vertex P can also carry the attractive force to the vertex V_0, if it has the same attribute as V_0. This is a major difference between the FR-algorithm and our new graph clustering technique. We rely on the similarities between vertices' attributes to obtain graph clustering. This force is defined as $f_c=T\times D^2/R$, where D is the distance between P and V_0, T is a constant determining the clustering speed. Thus the overall force at one point can be gained by:

$$F=w_1\times\sum f_a+w_2\times\sum f_r+w_3\times f_c \quad (1)$$

where w_i ($i=1$, 2, and 3) is the weight and $w_1+w_2+w_3=1.0$. The forces on other vertices are calculated in a similar manner. After the forces of all the vertices are computed, the locations of the vertices are updated. This process repeats until a stable state is reached when a clustered graph is obtained. Our experiments show that the settings of $w_1=0.2$, $w_2=0.2$, $w_3=0.6$, $T=2.0$ are appropriate for most graphs and 100 iterations are sufficient to obtain reasonable clustering results. The pseudo-code in Algorithm 1 provides further insight into the workings of our graph clustering with attribute similarities. Each iteration the algorithm computes O(E) attractive forces, O(V^2) repulsive forces and O(V) clustering forces where E and V respectively represents the number of edges and nodes.

Algorithm 1. Graph clustering with attribute similarities

Input: $G=(V,E)$: for each $v\in V$ has an attribute value. In total N discrete attribute values distribute on V. A position P_i ($i=1, 2\ldots N$) is set as the center of the cluster representing one attribute value.

```
for i:=1 to iterations do begin
  for v in V do begin
    for u in V and u ≠ v do begin
      calculate repulsive force f_r between u and v;
    End
    calculate attribute similarities force f_c
    between v and P_j (v and P_j have the same attribute value);
  End
  for e in E do begin
      calculate attractive force f_a between u and v;
      (u and v are the two endpoints of e)
  End
  for v in V do begin
    calculate overall force F=w1∑f_a+w2∑f_r+ w3∑f_c;
    update the position;
  End
End
```

3.2. Connections based on electric field

To properly visualize retweeting connections without node/link collision, we use a physical model to wire the connections. In a typical representation of a static electricity field, an arrowed line begins with a positive charge and ends with the negative charge. Such a line can be considered a message sent by a user at the beginning and retweeted by another user. So we can use this property to simulate the connections between users. Straight lines are used to represent the connections of adjacent vertices in the graph.

As shown in Fig. 3, the gray dots represent users. The user C_0 sends a message in microblog which is retweeted by N users such as C_1, C_2 and C_3. Here we regard C_0 as a positive charge and treat N other users as negative charges while setting $2\pi/N$ to θ. As a result, the retweeting connections between users are transformed into arrowed lines in an electric field. The linking process will be explained as follows.

Having determined the value of L, i.e. the length of one of the blue lines, a line at an angle of θ from the dot C_0 is drawn. Here each line at each time step has a uniform length of L. Before drawing the next line, the direction of the line should be determined. Assume P is one end point of this line just drew; all the charges will exert the Coulomb forces on it. Then P is regarded as a unit positive charge and q is its quantity of electricity. The same charges are repellent while opposite charges attract. N negative charges attract the end point and the positive charge repels it. So that the average force can be calculated as follows:

$$F_{\text{average}}=\sum_{i=1}^{N}k\frac{Q_i'q}{R_i^2}+k\frac{Qq}{R^2} \quad (2)$$

where k is a constant, Q and Q' represent the amounts of the positive charge and the negative charge respectively, and R is the distance between the end point and the charge. The direction of the average force would decide where the next line starts. In addition, each end point should be tested to determine whether it is beyond the boundary or arrives at a negative charge. If an end point arrives at a negative charge, the current electric field line is just what we need to present the connection. If the end point is beyond the boundary, the electric field line will be

deleted and not shown. Next, we should draw a new line from the dot C_0 at an angle of $2*\theta$. Then, other points can be connected in the same way.

However, some lines in the electric field are likely to be captured by the relatively close negative charges before reaching their destination charges, which result in more than one or no line between the adjacent vertices. In order to avoid this, the farther the negative charge is from the positive charge, the bigger the electricity value is set. The electricity value of each negative charge is defined as $Q' = md$, where m is a constant and d is the distance between the positive charge and the negative charge. During the calculation process, if the last end point of the line is coincident with a negative charge, this charge can be ignored in the later calculation process.

The layouts of edges from other nodes are all computed in the same way. The pseudo-code of our connections based on electric fields is summarized in Algorithm 2 below. The complexity of the whole progress is respectively $O(E^2/(VL))$ at best and $O(E^2/L)$ at the worst.

Algorithm 2. Connections based on electric fields

```
Input: G=(V, E), the positions of all nodes have been set;
       L is set as the step length in drawing edges.
for v in V do begin
       count the number of edges from v (denoted as N);
       treat N linked nodes(u1, u2 ...uN) of v as negative charges
       and v as positive charge with the electricity value Q;
       set the electricity value of the negative charge ui as Q'=md
       (m is a constant and d is the distance between ui and v);
       for i:=1 to N do begin
          draw a line l of length L at an angle of θ=i×2π/N from v;
          while(l reaches no linked nodes of v)
             calculate the direction of the next line l' from P as
                   F=∑kQ'q/Ri^2+kQq/R^2
             (P is the end point of l, Ri is the distance between ui and P
              and R is between v and P);
             draw the line l' of length L from P and set l=l';
          End
          set the electricity value of uj l reaches as 0;
       End
End
```

3.3. Prediction models on retweeting

Finally, to be able to predict on the number of future retweets based on the available retweeting data, we investigate diffusion models on their suitability to various retweeting patterns.

Different topics may be characterized by different numbers of retweets over time. Our observation shows that a hot topic is typically retweeted by a large number of users in a short period of time, and then the number gradually reduces over time.We use three diffusion models derived based on the statistical analysis of our large data sources (Tweets from about 200,000 users, which will be introduced in Section 5.1 in detail), shown in Fig. 4, denoted as A, B, and C models. We explored the retweeting data on a large number of topics and found about 60% topics showed the logarithmic normal distribution on the number of retweets over time (that is Model A). About 20% topics accorded the Model B and the left topics are mixed and disorderly as Model C.

Based on these diffusion models we propose a model-matching method to predict the future number of retweets.

Suppose we have the existing data $(x_1, y_1)\,(x_2, y_2)\ldots(x_n, y_n)$ where x_i represents a particular time point and y_i denotes the number of retweets during the period between x_i and x_{i-1}. The interval, denoted as t, between two adjacent time points is a parameter determined by the number of retweets, the retweeting time and the popularity of the topic. By analyzing a given set of data, we could estimate the diffusion model for the retweeting of a topic and predict the number of retweets y_{n+1} between x_n and x_{n+1}.

Model A: As the general model, this model shows the logarithmic normal distribution, as shown in Fig. 5. The equation of logarithmic normal distribution is as follows:

$$y = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x-\mu)^2/2\sigma^2} \tag{3}$$

where the fitted parameters $\hat{\mu}$ and $\hat{\sigma}$ are from the existing data. We can further obtain a new series of fitted values $y'_1, y'_2 \ldots y'_n$. By calculating the fitting precision of the fitted and actual values, we can estimate the accuracy of the fitness. The fitting precision denoted by R can be computed by:

$$R^2 = 1 - \left[\frac{\sum(y_i - y'_i)}{\sum(y_i - \bar{y})}\right]^2$$

where $\bar{y} = \sum_{i=1}^{n} y_i/n$. The closer R is to 1, the better the retweeting pattern matches this model. Assuming T_A is the threshold of Model A ($0.5 < T_A < 1.0$), if $R > T_A$, the retweeting pattern and Model A matches successfully. Insertingx_{n+1}, $\hat{\mu}$and $\hat{\sigma}$in Eq. (3), we obtain y_{n+1} as the predicted number of retweets between x_n and x_{n+1}. When R is smaller than T_A, the retweeting pattern does not match Model A.

Model B: A stable period is the typical feature of Model B. When testing whether the retweeting of a topic matches Model B, we calculate the mean ε and square deviation Φ of the recent five data items as follows.

$$\varepsilon = \frac{\sum_{i=n-4}^{n} y_i}{5} \qquad \phi = \sqrt{\sum_{i=n-4}^{n} (y_i - \varepsilon)^2/5} \tag{4}$$

If Φ is close to 0, the retweeting pattern tends to be stable. Assuming T_B is the threshold of a pattern, when $\Phi < T_B$, the pattern is considered to match Model B successfully. An arbitrary value based on $N(\varepsilon, \Phi)$ can be assigned to y_{n+1} as the number of retweets between x_i and x_{i+1}. If $\Phi > T_B$, the retweeting pattern does not match Model B.

Model C: The retweeting pattern in Model C is the most complex, making a prediction based on all the given data impossible. We therefore predict the next time step simply based on the most recent data. We can build a cubic curve:

$$y = ax^3 + bx^2 + cx + d \tag{5}$$

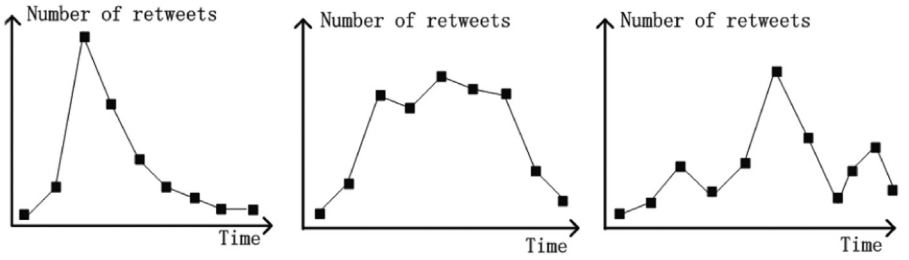

Fig. 4. Three diffusion models.

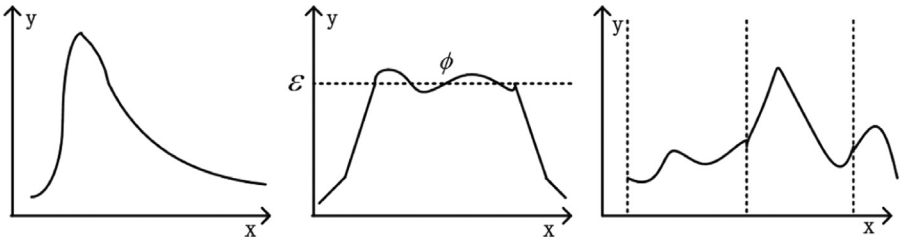

Fig. 5. Fitting curves coinciding with the three diffusion models in Fig. 4.

where the fitted parameters a, b, c, and d represent the data within the recent five time steps using the least square principles [17]:

$$a = \frac{y_n - y_{n-4} - 2y_{n-1} + 2y_{n-3}}{12}$$

$$b = \frac{-3y_{n-2} - y_{n-3} - y_{n-1} + 2.5y_{n-4} + 2.5y_n}{18}$$

$$c = \frac{8y_{n-1} - 8y_{n-3} - y_n + y_{n-4}}{12}$$

$$d = \frac{6y_{n-2} - y_{n-4} - y_n + 4y_{n-1} + 4y_{n-3}}{12}$$

It is also important to note that the above equations of the fitted parameters are derived from the least square method by assigning -2, -1, 0, 1, 2 to x_{n-4}, x_{n-3}, x_{n-2}, x_{n-1}, x_n respectively. Therefore, by assigning a value of 3 to x_{n+1}, and inserting x_{n+1}, a, b, c and d in Eq. (5), we can predict the number of retweets y_{n+1} between x_n and x_{n+1}.

Generally the retweeting pattern for each topic could match one of the three models discussed above. Models A and B can be used to study topics that have typical characteristics in the number of retweets over time. There are in fact also many other special retweeting patterns. Yet Models A and B are the most common and representative. In order to simplify the prediction process, the topics whose retweeting patterns do not match Models A and B are all analyzed using the Model C. Predicting the number of retweets on a topic using the Model C is simply based on the latest data. The experiments on some topics demonstrate the effectiveness of our method in Section 5.2.

4. Interactive visualizations for microblog retweeting

Based on the above data analysis approaches, three visualization views are designed in MV^2 as follows to show different aspects of the retweeting data. All the three views could be interactively explored. Any change in one view would automatically propagate into and update other views.

4.1. Location analysis with virtual map

The first view is on the geographical locations of retweets, adapted from Facebook. Mapping Facebook Friendships [18] developed by Facebook intern Paul Butler in 2010, maps the connections of 10 million Facebook users spread all over the world. Our virtual map is based on the similar idea, and maps the retweeting connections of 200,000 microblog users spread across China. Using a map of China as the background can directly show the geographical layouts of network nodes. But the map details may clutter the layout of the connections. We therefore use a simplified map with the major cities represented as points.

We match users with the corresponding coordinates of each city, and then connect them dynamically in bundled lines denoting retweeting relationships over a period of time by the divided edge bundling method [8]. Lines fade from red (source) to blue (target) to indicate the retweeting direction (see Fig. 6). As the number of retweets in the city is growing, the color of the corresponding point is becoming brighter. On the other hand, the corresponding connection line will become wider, when the retweet frequency between two cities is increasing. It will help identifying these cities which play an important role in retweeting. We can further explore whether there is a special relationship between the microblog users of two cities.

4.2. Clustered graph

The second view is to visualize retweeting connections clustered on users' occupations. Each vertex in a clustered graph represents a user. The vertices are initially positioned randomly. Then we update all vertices' coordinates based on users' occupations by clustering (see Section 3.1). The updates of positions are animated dynamically in the visualization.

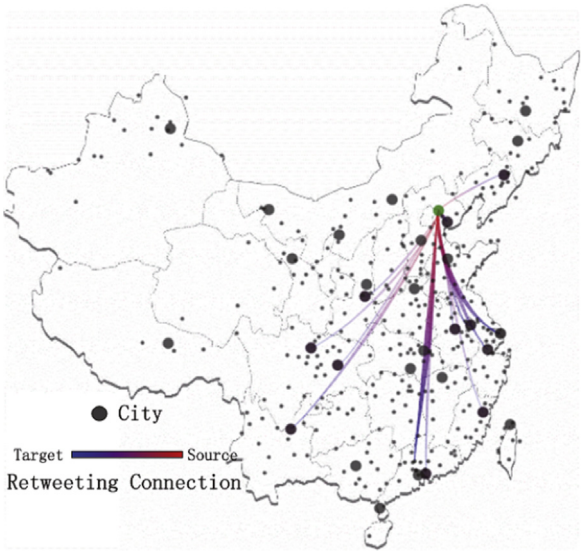

Fig. 6. Virtual Map mapping retweeting connections among major Chinese cities. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

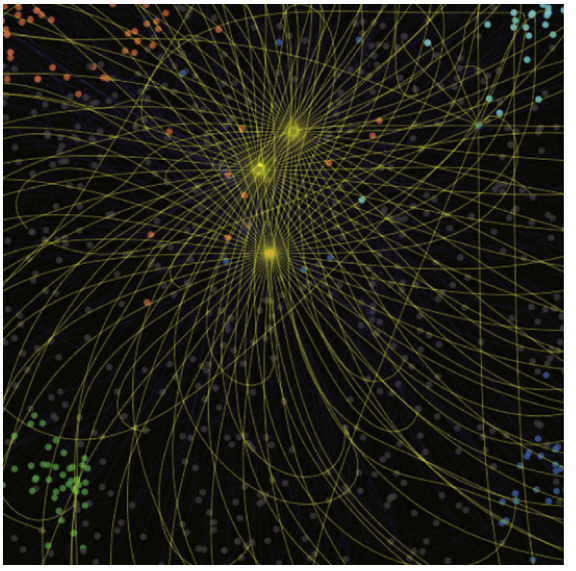

Fig. 8. Layout of connections. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

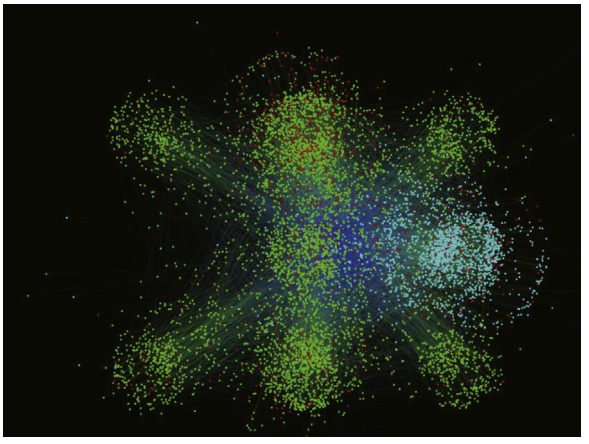

Fig. 7. An example clustered graph based on attributes similarities.

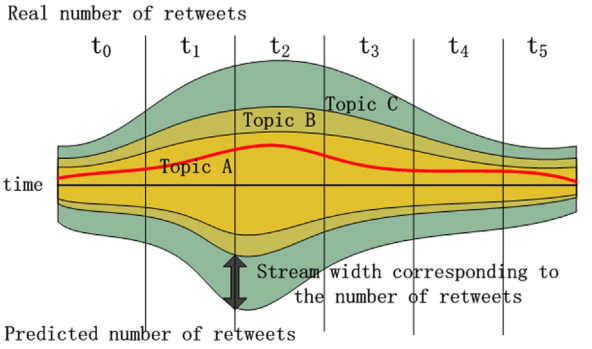

Fig. 9. Retweet Streams illustrating evolution of three topics. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

When the animation process ends, we can clearly see several groups within the graph, as shown in Fig. 7. Here eight groups are generated in this example. The group size is relative to the number of participants of a particular occupation.

Then the retweeting process begins, with yellow lines denoting retweeting connections from one to another shown dynamically over time. The graph layout is generated using the force-directed drawing method (see Section 3.2), as shown in Fig. 8. Using the keyboard or mouse, we can move the graph and zoom in or out to view the details. By clicking on any interesting vertex, we can see the corresponding user's account name and comments.

4.3. Retweet Streams

The third view, Retweet Streams is used to encode both predicted and actual numbers of retweets over time. The streams are stacked, giving the impression of a river consisting of multiple streams representing different topics.

The *x*-axis of the visualization is used as the time dimension. The *y*-axis encodes the relative number of retweets, while each stream band corresponds to a topic. We divide the window into two parts, the upper part represents the actual number of retweets and the *y*-axis of the lower part mirrors that of the upper part representing the predicted number of retweets. Through the symmetry between the two parts, we can know the correctness of our prediction based on the model-matching method discussed in Section 3.3.

Different colors are used to encode different topics. As shown in Fig. 9, there are three topics expressed in various tones of green. The topic is placed in the center. If the original posters take up large shares of the total retweets, we will visualize their entire retweets with red flows. The height and width of each flow are both determined by the number of retweets in the corresponding time point.

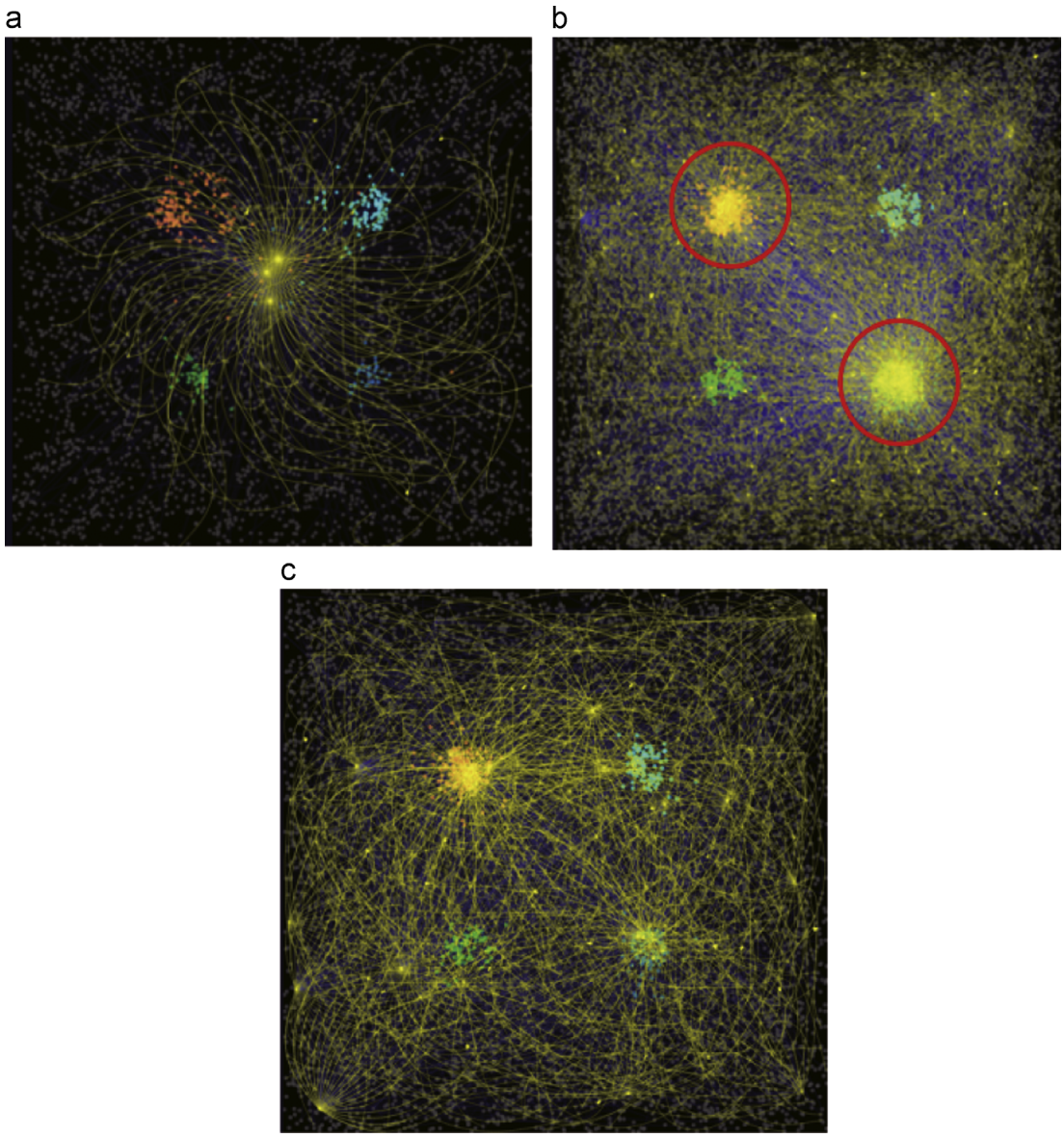

Fig. 10. Clustered Graphs for the three topics. (a) Topic A, (b) Topic B and (c) Topic C. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

5. Real-life case studies

5.1. Data collection and analysis

Sina Microblog is our empirical data source, which has the largest microblog community in China with over 100 million users. Tweets from 200,000 users on five popular social issues in 2012 were crawled for our empirical study. Having explored the retweeting data, we selected about 10%, i.e. 20,000 typical users. Original messages are usually posted by a small proportion of users. Majority users retweet messages and add comments.

The data we collected include retweeting connections, users' messages or comments, and users' profile information, including gender, address, membership level and occupation. We divide the users' occupations into five broad categories: celebrities, public figures, journalists, academics and others. The first four categories are encoded in different colors: cyan, blue, orange and green. The gray vertices represent others and those whose occupation information is missing. The five social topics attract much of the public's attention over a long period of time. Based on the average retweeting time which is typically in hours, we make 6 h as a unit of time, and count the number of retweets for each topic.

We applied MV^2 to three different topics which are: Contributing Money by Retweeting, Li Zhuang's Case, Deng's Attempted Assassination of a Government Official. They are denoted by Topics A, B and C respectively.

Briefly, Topic A is about donating money on microblogging to help a child who was badly scalded. A real-estate manager promised that anyone retweeted once about this topic, he would donate RMB 1. At the end, the manager donated RMB 10,000 and the child received about RMB 130,000 from ordinary donors. Topic B is the story of a lawyer committing perjury to harbor criminals. Mr. Li Zhuang, a local lawyer, had defended a dozen of criminal suspects by making false statements and finally the criminal suspects were all acquitted. This event caused much controversy and debate over the Chinese legal system, and also attracted academic attention. Topic C is on a waitress named Miss Deng Yujiao who stabbed a government official on self-defense after being harassed by the official. The local police took arrested Miss Deng on suspicion of "intentional homicide". This event caused a nationwide discussion on the Web. The public expressed sympathy for Miss Deng condemned the official.

5.2. Data visualization

As shown in Fig. 10, comparing the three topics, Topic A has the fewest lines, demonstrating that it has the least number of retweets. Meanwhile, the low intensity of four clustered groups shows weak relationships between retweeting connections and occupations. But there are three highly-popular users shown as the bright spots representing their activities. Then, Topic B has the most number of retweets, showing the largest number of lines. Two groups (public figures and journalists denoted as red circles as shown in Fig. 10b) are highlighted, showing that they made a significant impact on the propagation of the topic. By contrast, only one occupation (journalists) is highly active in Topic C.

The time distributions of retweeting of the three topics appear different in Fig. 11. The number of retweets on Topic A was rising rapidly in the first 6 h. It then entered a stable period for the next 18 h. In the final phase, the number gradually reduced to a low value. Meanwhile, only one original user made up a large share of the total retweets. The retweeting pattern on Topic B matches well with the Model A as discussed in Section 3.3. The hot tweets were retweeted by a large number of users in the first 18 h, and then the number of retweets gradually reduced during the next 30 h. In contrast, there were two peaks in the number of retweets for Topic C, which happened in the 48th and 104th h.

As discussed in Section 4.2, we divide the window of Retweet Streams into two parts. The upper part represents the actual number of retweets and the lower one indicates the predicted number of retweets. It can be seen from Fig. 11 that the simulation for Topic B produces the best result. Because the Model A matches well with retweeting patterns for most hot topics, we can make a prediction using the Model A on the number of retweets. We can see that the simulation for Topic A has a time lag to some extent. The simulations of that match Model B have this common problem. It is difficult for us to know when the number of retweets enters a declining period from the stable period. Finally, the simulation result for Topic C which matches the Model C is the worst, as the retweeting patterns on those topics are the most complex. We predict the next time step simply according to the latest data.

Topic A has only two active cities (Beijing and Guangzhou), illustrated in Fig. 12a. In other words, most users in other cities retweeted the original message posted by users in Beijing and Guangzhou. Besides, the low density of the lines indicates that Topic A does not cause much social concern. The results of Topic B are seen as clear evidence that

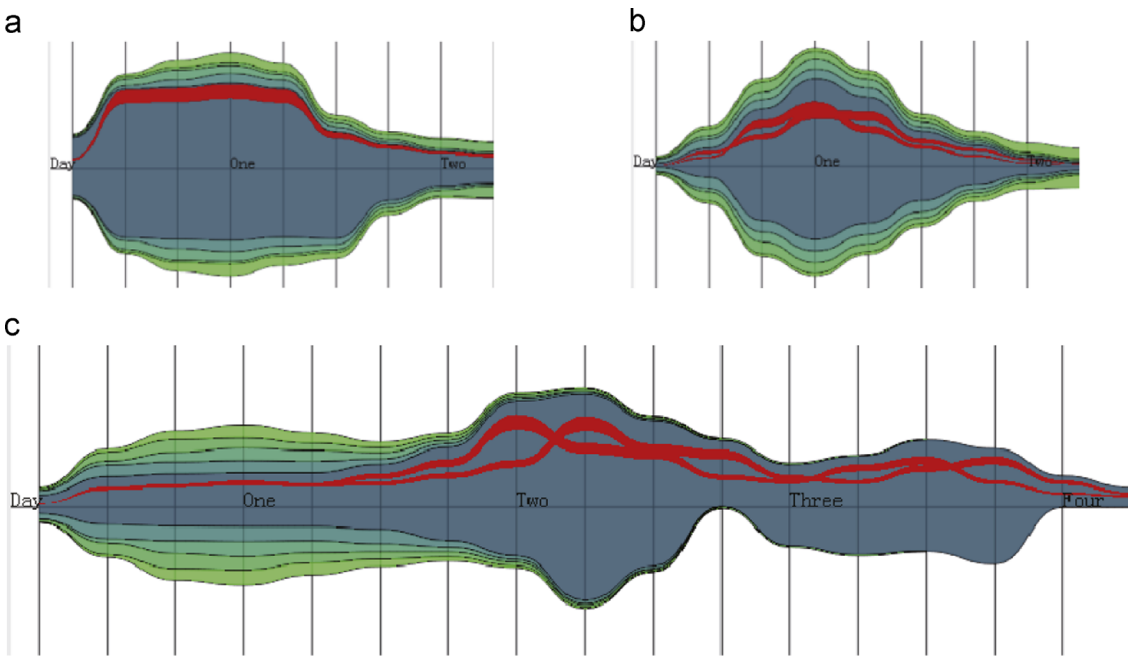

Fig. 11. Retweet Streams for the three topics. (a) Topic A, (b) Topic B and (c) Topic C.

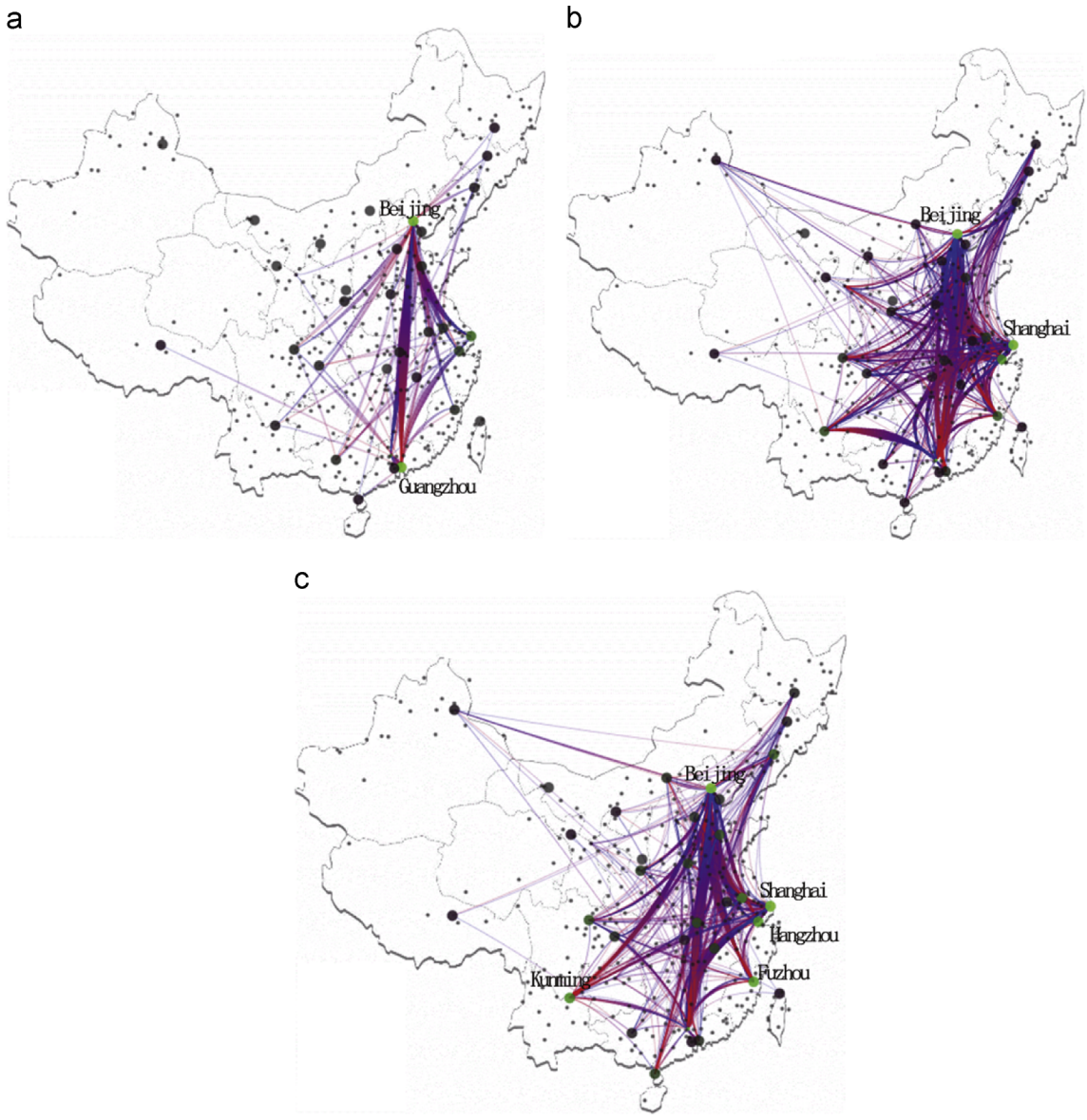

Fig. 12. Virtual Map for the three topics. (a) Topic A, (b) Topic B and (c) Topic C.

the two most active cities are Shanghai and Beijing. The density of the lines is much higher than Topic A, which shows Topic B produced a nationwide discussion. We can see that topic C has the largest number of active cities in the three topics. It does suggest that the original messages were more widely distributed. Beijing, Shanghai and Kunming are the three most important active cities. Many users chose the original messages from Beijing to retweet, since Kunming's point of view is less provocative than Beijing's. Because the density of the lines is about average, it indicates that comments were not highly enthusiastic.

5.3. *Retweeting management and evaluation*

As shown in Fig. 13, our system architecture consists of three primary components. First, in the raw data module, micro-blog data are collected from the Sina micro-blog API. These raw data are cleaned and stored in a database to support temporal explorations. The data gathering mode provides a monitoring approach that allows users to monitor current events. Meanwhile, users can query the historical data. The layout module supports efficient layout methods discussed in Section 3 that transform and render the raw data into various visualization forms. The parameters (e.g. color schemes and appearance of visual elements) used to render the visualization are all customizable to suit different user communities, such as government decision-makers and social science researchers. Finally, user-friendly interactions are also supported. Users' inputs feed back to the rendering and data module to enable data exploration. Using the keyboard, one can choose any interesting topic and manipulate the virtual map and the clustered graph or zoom in/out

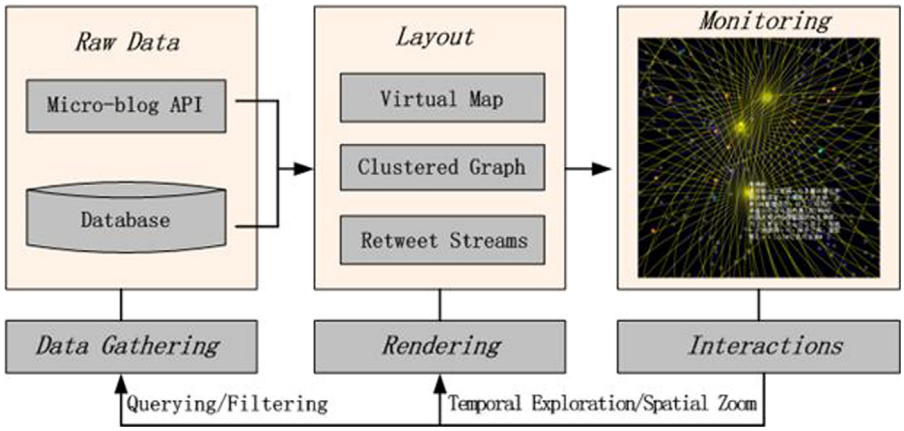

Fig. 13. Framework of MV^2 system.

Table 1
Comparison of MV^2 with other related approaches.

Related approaches	Time-continuity	Geo-distribution	Sentiment analysis	Prediction	Sample size	Retweeting relationships
Visual backchannel [14]	Strong	No	No	No	50,680	No
Whisper [23]	Strong	Yes	Yes	No	1594	Yes
Pixel sentiment calendar and Geo Map [16]	Weak	Yes	Yes	No	59,614	No
MV^2	Strong	Yes	No	Yes	200,000	Yes

to view the details. By clicking on any interesting user node in the clustered graph, we can see the corresponding microblog user's information such as account name, comments and address. The visualization reveals that many tweets become popular after being retweeted by an important opinion leader. It is useful to trace important news and events by illustrating the detailed diffusion patterns of "when did who retweet whom".

As summarized in Table 1, ecompared with other approaches, such as Visual Backchannel [14], MV^2 visualizes retweeting relationships and clusters. It is able to predict the number of retweets over time and visualize such a trend in the social networks, although it does not perform sentiment analysis as Whisper [23] and Pixel Sentiment Calendar and Geo Map [16] do. We conducted a questionnaire guided by the following questions:

(1) Is the topic of regional or nationwide interest?
(2) When and where does the topic attract attention? By which city or which occupation?
(3) How do different cities or occupations respond to the topic?
(4) Are there any opinion leaders discussing this topic? When and how does that happen?
(5) How does information propagate differently over time or cities/occupations for different topics?
(6) How does the system help predict the number of retweets over time for different topics?

We then explain the purpose and features of MV^2 to 50 users. After they explored the micro-blog data about the three topics above using MV^2, their responses to the questionnaire and comments on the system were recorded in detail. We summarized the users' feedback as follows. All the users agreed that it is easy to see how a topic becomes popular. They believed that MV^2 could clearly show the trend of a topic and the types of participants. Using MV^2, they could easily find which topics worth paying attention. The system showed a multi-layer bursty topic so similar to what exactly happened. Particularly noteworthy was the forecast based on real data and simulated results that can ensure the accuracy of simulation.

6. Conclusion

In this paper, we have presented MV^2, a visualization tool for analyzing large-scale retweeting behaviors on different topics. Compared with previous research, our work is unique in two ways. First, it enables better understanding about retweeting evolutions within cities and among different occupations. Using the proposed visualization, we can discover highly-popular or active users, cities and occupations and further explore the potential relationships among them. Second, we have proposed a new method for predicting the number of retweets. Three diffusion models are proposed to match the retweeting patterns.

How to find a good balance between the clustering of vertices and the graph layout is still a challenge. Connections based on electric field reduce the number of edge crossings only to a certain extent. In addition, we have only explored the retweeting behaviors on a single topic.

There is a need for much more research on the influence of retweeting behaviors on multiple different topics.

Acknowledgments

This paper was partially supported by the Natural Science Foundation of China under Grant no. 61272199, Doctoral Fund of Ministry of Education of China under Grant no. 20130076110008, and National High-tech R&D Program of China (863 Program) under Grant no. 2012AA011003, the Innovation Program of the Shanghai Municipal Education Commission under Grant no. 12ZZ042, and the Shanghai Collaborative Innovation Center of Trustworthy Software for Internet of Things under Grant no. ZF1213. The authors would like to thank W. Qian, H. Ma, and Q. Zhang for providing the data on Sina microblog.

References

[1] B. Suh, L. Hong, P. Pirolli, E.H. Chi, Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network, in: Proceedings of the Second IEEE International Conference on Social Computing, 2010, pp. 177–184.
[2] J. Xie, C. Zhang, M. Wu, Modeling microblogging communication based on human dynamics, in: Proceedings of the Eighth International Conference on Fuzzy Systems and Knowledge Discovery, 2011, pp. 2290–2294.
[3] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, Z. Su, Understanding retweeting behaviors in social networks, Proceedings of the 19th ACM international conference on Information and knowledge management. ACM (2010) 1633–1636.
[4] P. Eades, A heuristic for graph drawing, Congr. Numerantium 42 (1984) 149–160.
[5] T. Kamada, S. Kawai, An algorithm for drawing general undirected-graphs, Inf. Process. Lett. 31 (1989) 7–15.
[6] R. Davidson, D. Harel, Drawing graphics nicely using simulated annealing, ACM Trans. Gr. 15 (4) (1996) 301–331.
[7] T. Fruchterman, E. Reingold, Graph drawing by force-directed placement, Softw.: Pract. Exp. 21 (11) (1991) 1129–1164.
[8] D. Selassie, B. Heller, J. Heer, Divided edge bundling for directional network data, IEEE Trans. Vis. Comput. Gr. 17 (12) (2011) 2354–2363.
[9] J. Moody, D. McFarland, S. Bender-deMoll, Dynamic network visualization, Am. J. Sociol. 110 (4) (2005) 1206–1247.
[10] K. Durant, A. McCray, C. Safran, Modeling the temporal evolution of an online cancer forum, in: Proceedings of the 1st ACM International Health Informatics Symposium, 2010, pp. 356–365.
[11] S. Bender-deMoll, D. McFarland, The art and science of dynamic network visualization, J. Soc. Struct. 7 (2) (2006) 1–46.
[12] P.A. Gloor, J. Krauss, S. Nann, K. Fischbach, D. Schoder, Web science 2.0: identifying trends through semantic social network analysis, Comput. Sci.Eng. 4 (2009) 215–222.
[13] L. Byron, M. Wattenberg, Stacked graphs-geometry aesthetics, IEEE Trans. Vis. Comput. Gr. 14 (6) (2008) 1245–1252.
[14] M. Dork, D. Gruen, C. Williamson, S. Carpendale, A visual backchannel for large-scale events, IEEE Trans. Vis. Comput. Gr. 16 (6) (2010) 1129–1138.
[15] N. Diakopoulos, M. Naaman, F. Kivran-Swaine, Diamonds in the rough: social media visual analytics for journalistic inquiry, in: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, 2010, pp. 115–122.
[16] M. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D.A. Keim, L.-E. Haug, M.-C. Hsu, Visual sentiment analysis on twitter data streams, in: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, 2011, pp. 277–278.
[17] S. Roychowdhury, Fuzzy curve fitting using least square principles, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 1988, 4, pp. 4022–4027.
[18] P. Butler, Mapping Facebook Friendships. 2010. ⟨http://on.fb.me/hy6dmb⟩.
[19] A. Marcus, M. Bernstein, O. Badar, D. Karger, S. Madden, R. Miller, Twitinfo: aggregating and visualizing microblogs for event exploration, in: Proceedings of the ACM CHI, 2011, pp. 227–236.
[20] E.R. Gansner, Y. Hu, S.C. North, Visualizing streaming text data with dynamic graphs and maps, in: Proceedings of the Proceedings of Graph Drawing, 2012, pp. 439–450.
[21] D. Archambault, D. Greene, P. Cunningham, N. Hurley, ThemeCrowds: multiresolution summaries of twitter usage, in: Proceedings of the Proceedings of the Workshop on Search and Mining User-generated Contents, 2011, pp. 1–20.
[22] Google+Ripples: revealing how posts are shared over time, in: Visualization Blog on Information Aesthetics, 31 October 2011.
[23] N. Cao, Y.R. Lin, X. Sun, D. Lazer, S. Liu, H. Qu, Whisper: Tracing the spatiotemporal process of information diffusion in real time, IEEE Trans. Vis. Comput. Gr. 18 (12) (2012) 2649–2658.