Telematics and Informatics 34 (2017) 755–764

Contents lists available at ScienceDirect

Telematics and Informatics

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

Interdisciplinary study on popularity prediction of social classified hot online events in China

Tieying Liu [a,*], Yang Zhong [a], Kai Chen [b]

[a] School of International and Public Affairs, Shanghai Jiao Tong University, Shanghai 200230, China
[b] Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200230, China

ARTICLE INFO

Article history:
Received 30 September 2015
Received in revised form 6 April 2016
Accepted 5 May 2016
Available online 6 July 2016

Keywords:
Popularity prediction
Online hot event
Political event
Social event
Non-public event

ABSTRACT

We offer an interdisciplinary study of computer science and social science, analyzing behavior surrounding three types of online events: political events, social events, and non-public events. Based on the intrinsic characteristics of the three event types, this paper creates an effective method to predict such events. We continuously followed and recorded data every 10 min for 10 months from September 14, 2012 to July 11, 2013, and collected over 14 million "hot" posts from Sina Weibo, the largest microblogging provider in China. After removing spammers and noises, we developed a database of 4180 hot online events and 7,761,395 threads. We found that people's online behavior regarding event types varies in terms of follow-up statistics and the predictability of events. The Chinese are, typically, quite concerned with social affairs that relate most closely to their personal interests and preferences. People tend to cluster around political events more often than social events and non-public events. This is demonstrated by an algorithm embedded with a clustering growth pattern of events, which predicts the popularity of online political events above others. The statistical findings are justified by Habermas' public sphere theory and the theory of vertical/horizontal collectivism/individualism. This research provides an interesting piece of computational social science work to assist in the analysis of incentives concerning China's collective events.

1. Introduction

Nowadays, China is home to the world's largest number of ‘‘netizens," or people who are active in online communities. The 31st Statistical Report on Internet Development in China, issued by the China Internet Network Information Center, reported that the total netizen population in China was 688 million in December 2015, with an Internet penetration rate of 50.3% of the total population. The popularity of Internet enhances Chinese's participation in public affairs. Understanding people's attitude and behavior model toward political affairs and social affairs online is becoming another way for us to understand Chinese society. The Internet plays a special role in China: it is an important channel for Chinese to participate in social affairs. The Internet has also expanded the freedom of information in China. Research into online hot events diffusion will help us understand Chinese preferences and concerns regarding public participation.

Online events reflect real-world situations of China economically and politically to some extent. In recent years, more and more collective events have taken place in China, which may compose potential threat to the social stability. As the world's

[*] Corresponding author.
E-mail addresses: tyliu@sjtu.edu.cn (T. Liu), zhongyang@sjtu.edu.cn (Y. Zhong), kchen@sjtu.edu.cn (K. Chen).

http://dx.doi.org/10.1016/j.tele.2016.05.022
0736-5853/© 2016 Elsevier Ltd. All rights reserved.

second largest economy, questions have been raised for many years concerning China's social and political stability. On the one hand, China's economic reform over the past three decades is considered an obvious success. China's economic system has been transformed from a command economy to an expanding market economy (Lewis and Litai, 2003). Furthermore, its fast economic growth has ensured extensive support and social prosperity. On the other hand, rapid economic growth has resulted in social disruption, pervasive corruption, and social inequality (Cui et al., 2014; Ho, 2014). The complex change in China's social strata brought about by economic reform has intensified unrest and insecurity (Knight, 2013). In contrast, China's political system has remained unchanged in recent years, but mounting political and social problems have introduced volatility into society (Shirk, 2007). Large-scale online collective behaviors occur frequently in China (Sullivan, 2014; Yang, 2013). Research on the patterns of online collective behaviors with massive data opens more accessible way to observe Chinese interests and concerns regarding public affairs than conventional (usually offline) social questionnaire methodologies with sparse data.

In this paper, we conduct an interdisciplinary study of computer science and social science by investigating the following questions: 1) Are the behavior modes of netizens engaged in online hot political events or social events characterized by any special features? 2) Can we better predict the popularity of various online collective events based on netizens' behaviors?

We offer an interdisciplinary research framework including Internet data collection, algorithmic popularity prediction of online hot events, and the related explanations about experimental findings from the viewpoints of public sphere (Habermas, 1991), individualism/collectivism and some offline empirical researches. Our empirical study of China's online hot events is based on more than 14 million posts collected from Sina Weibo between September 14, 2012 and July 11, 2013. Controversial opinions regarding China's current regime have inspired us to classify the hot events into three categories: political public events, social public events and non-public events, as based on a bag-of-words model used in social science (Dumais, 2004). Such classifications can help us observe the root reason for China's hot events and better understand people's collective behaviors and intentions.

To compare and predict the popularity of the three types of events, we used three algorithms: a linear algorithm, correlation-based algorithm and a modified state transition-based algorithm. For online political events, the modified state transition-based algorithm visibly improves the prediction accuracy compared with the other two algorithms. We found that the growth pattern for political events is more regular and predicable than the other two types of online events and people are more likely to cluster and converge on online political events.

The prediction results provide an interesting piece of computational social science work to assist in the analysis of incentives concerning China's collective events. We discuss their far-reaching implications for Chinese politics and predict for China's social stability justified by public sphere (Habermas, 1991), horizontal/vertical individualism/collectivism, and some offline empirical studies and observation.

2. Literature review

About the impact of social media on public sphere, there is a lot of debate. Generally social media offers increasing opportunities for political communication and enable democratic capacities for political discussion within the virtual public sphere (Loader and Mercea, 2011). Habermas initially defined public sphere as "a domain of social life in which such a thing as public opinion can be formed." Access to this domain is "open in principle to all citizens" who may "assemble and unite freely, and express and publicize their opinions freely." He later recognized the internal dynamics of the public sphere, the possibility of multiple public spheres, as well as the conflicts and interactions among them (Habermas, 1991). Huang argues that Habermas binary opposition between state and society is inappropriate in China (Huang, 1993). He stated there is a third space in between state and society, in which both participate. The mass media constitute a source of power (Jarren and Donges, 2002). The dynamics of mass communication are driven by the power of the media to select, and shape the presentation of, messages and by the strategic use of political and social power to influence the agendas as well as the triggering and framing of public issues (Habermas, 2006). The rapid diffusion of new media has transformed the supply of information. There is a much wider range of media choices on offer, providing much greater variability in the content of available information. This means that something approaching information "stratamentation" (stratification and fragmentation at the same time) is going on. More people are drifting away from the mainstream media (Bennett and Iyengar, 2008). But on the new media area, unlike that in the US, Chinese government has large size and sophisticated censorship (King et al., 2012). Chinese Internet users have turned to the Internet for public expression and political activism even as the government is stepping up control (Barme and Davies, 2005). This offers Internet research in China an important role in public sphere and political communication.

As one aspect of political philosophy, the theory of individualism and collectivism influences people's attitude on public affairs and their behavior in public sphere. Hofstede first measured individualism and collectivism across cultures, the original two-dimensional conceptualization has been a successful predictor of behavioral patterns (Triandis and Gelfand, 1998). Triandis highlighted that it is important to make the distinction between vertical and horizontal individualism and collectivism (Triandis, 1995). Thus making four types of dimensions, horizontal collectivism, vertical collectivism, horizontal individualism and vertical individualism. Horizontal collectivists merge with in-groups, but they do not feel subordinate to their in-groups. Vertical collectivists submit to the norms of their in-groups and are even willing to sacrifice their personal identities for their in-groups. Horizontal individualists are characterized by seeking individuality rather than distinctiveness.

Vertical individualists are especially concerned with comparing themselves with others they desire to win in all kinds of competitions (Chiou, 2001). The conception is now considered fundamental to the understanding of cultural values (Triandis, 2004). The perception about collectivism and individualism in China's context is relatively intricate. China has historically been regarded as an interesting and complicated case of collectivism (Triandis, 1995). Traditional Chinese culture may be inclined toward verticality, but the State advocates horizontal themes. China may be considered a horizontal collective culture, influenced by the historical political and economic environment that emphasized egalitarian and group centered values. Sivadas' measurement results show China is vertical individual country (Sivadas et al., 2008).

Massive online data and advanced analytics enable us to extract the patterns of online collective behaviors in virtual public sphere. Online data from social media captures online behavior of users who communicate or interact on a diversity of issues and topics. Research shows that user attention is allocated in a rather asymmetric way, with most content receiving only limited views and just a few receiving significant attention. Some research results also indicate that perceived usefulness of bloggers' recommendations and trust had significant influential effect on blog users' attitude toward and intention to online shopping (Hsu et al., 2013). Previous studies tend to focus on commercial applications (e.g., such as prediction of software download, online purchase, and advertisement hits) to predict attention levels (Applegate et al., 2010; Jansen et al., 2009; Liang, 2014). Using prediction in commercial areas is more about people's commercial behavior, thus generating profit models for businesses. However, it ignores social values necessary to understand people's behavior on public affairs and offers no benefit for social research. The complexity of China's society makes it hard to understand the society, especially its political communication. It motivates us to study social classified hot online events related to collective behaviors.

Some researches help predict people's behavior, for example, a recent study predict the potential responders of a routed online question (Liu and Jansen, 2014). They found that it is possible to predict online content (Szabo and Huberman, 2010). Online large-scale data analysis provides a channel to analyze people's intentions regarding public affairs. With the development of technology the Internet provides ordinary citizens with greater social resources and opportunities to expand their level of public participation, and enhances citizen engagement in public affairs (Zheng and Wu, 2005). The Internet opens a new window to observe people's interests and concerns regarding public affairs, especially in authoritarian states.

3. Data

3.1. Data collection and screening method

Our research is based on hot events from Sina Weibo. As the largest microblog platform in China, Sina Weibo has 69.4% of Chinese microblog users; people catch, share and diffuse information here (CNNIC, 2015).

Hot events refer to incidents attracting significant attention, and are quantitatively expressed as receiving numerous replies and follow-up threads. It is very possible that many users initiate their own microblogs with the same topic. Thus, we need to define hot events, collect all threads relating to these hot events, combine the threads of the same topic, analyze the attention received from one event, and predict other events based on the data we collect.

One method would be to first collect all the data from Sina Weibo, and then screen and recognize the hot events. However, collecting all Sina Weibo's data could be regarded as an attack on the website server. In addition, collecting all the threads from only one website does not consider online hot events as a whole, and may exclude some events that are hot but have generated very little discussion on Sina Weibo. We fully understand that the hot event data we collected using this method do not represent online hot events as a whole.

People use microblogs to obtain information and to discuss hot events. Generally, online hot events also draw attention in the microblog world. Therefore, the ranking of hot events by authoritative websites provide a good reference base. Such websites cover a large landscape of data and have their own unique algorithms to ensure credibility and reliability. We adopted real-time hot events ranking lists from well-known search engines in China as our microblog hot events topic sources, including Baidu Ranking (top.baidu.com), Souhu Ranking (top.sogou.com), Tenet Ranking (top.soso.com), and Sina Ranking (huati.weibo.com).

We caught the top-10 hot events every 10 min from the rankings of these websites' real-time updated lists. We searched the titles of these hot events in Google news (news.google.com) and Baidu news (news.baidu.com) and obtained more text information and key words using TF-IDF (Term Frequency–Inverse Document Frequency) (Salton and McGill, 1983). With the text information and key words, we searched and analyzed the data from Sina Weibo to obtain all posts concerning these topics.

We then followed the online data for 10 months and caught over 14 million threads. However, we know that there are many online spammers. To remove such noise we selected active users by calculating the users and their friends' activities and relationships. We computed the number of user i's posts $N_i^{(0)}$ and calculated the number of threads followed by i's friend $N_i^{(1)}$. Furthermore, U_i^n is a user set, $n = N_j^{(1)}$, which is the number of rebroadcasts from j's friends. Here, user j has rebroadcast user i's content at least once. We used the following formula to calculate each user's score:

$$s_i = \sum_n n|U_i^n| + N_i^{(1)}$$

where,

$N_i^{(0)}$ is the number of user i's posts and reposts;
$N_i^{(1)}$ is the number of user i's friend posts and reposts; and
$|U_i^n|$ is the cardinality of the set U_i^n

$$s_i = 0 \quad if\ N_i^{(0)} < T_c \quad (T_c = 3)$$

We ranked the users according to their scores (and eliminated low-scoring users) to get our informative users dataset. We found that our active users comprised 70%–80% of an event's total popularity. Therefore, we can predict the popularity of a hot event by monitoring the behavior of informative users. We screened off 253,565 users; the remaining 6,151,912 users' posts were included in our research dataset.

We used a 24-h study window. The popularity of the first 5 h was obtained and we used that data to predict and observe the popularity of the next 19 h. The prediction result was deemed to be accurate if it was between 20% higher or lower than the real figure.

3.2. Data classification standard

As mentioned above, we manually divided all events into three categories: political public affairs, social public affairs and non-public affairs. Public affairs can be defined in a number of ways. Based on the state theory, states originate from the transfer of public power, which carries three basic functions: legislation, administration, and jurisdiction. From this perspective, public affairs are concerned with state sovereignty and legitimacy. It is also defined as the management of social affairs concerned with a common interest. Furthermore, public affairs can be regarded as a personal experience in public activities (Wang, 2001).

Political public affairs involve political system, legislation, the maintenance of social order, public safety, sovereignty, diplomatic relationships, and derivative administrative issues such as bureaucratic administration, fiscal administration, and governmental internal control. Political public affairs have no direct relationship with public interest. People's concern regarding political affairs shows their attitudes, values, and beliefs about society and government. These behaviors are geared toward restoration of justice (Qiu et al., 2014). People's agreement on public affairs of this sort represents a fundamental support of the political system and governance. Therefore, political participation is related to the evaluation of how well political institutions conform to a person's sense of what is right (Muller and Jukam, 1977). Conflicts arising from political disagreement are hard to resolve because they concern basic values and beliefs. It is for this reason that it is difficult to reach agreement in collective events of a political nature.

Social public affairs relate to people's daily lives and interests; they can involve all members of society and have the greatest effect on people's interests. Social public affairs include education, science and technology, medical and health systems, and transportation (Wang, 2001); they influence people's interests, are interest driven and have no relation with personal values and beliefs. Because of the close relationship with public interest, people may pay greater attention to social public affairs, and may behave radically or emotionally to express their opinion or satisfaction. However, such conflicts do not stem from a fundamental value or belief, but from their temporary benefit requirement or emotional contagion. Once their interest is satisfied or outrage has reduced, the behavior will stop. Support for some social events may change very quickly once interests are harmed. Social events pose little risk to the political system or regime.

For the purposes of this study, all events not considered political public events or social public events are defined as non-public events. These events include points of interest that basically have no influence on the public, such as the private lives of pop stars, a high-class event, or humorous occasion.

4. Prediction methodologies and calculation

Various algorithms have been used to predict the popularity of hot events, including time series algorithm, rule-based classification algorithm, and Bayesian network algorithm (Dumais, 2004).

Time series algorithms are cost-effective methods to predict an event's popularity using less data. People's behaviors tend to emerge spontaneously under the influence of others (Onnela and Reed-Tsochas, 2010). Such phenomena can be explained by the herd effect online (Duan et al., 2005; Lee and Lee, 2012). We introduced an affinity propagation algorithm into the time series algorithm to help us improve the accuracy of the prediction results. Because people hold different attitudes and preferences regarding political events, social events, and non-public events, we assumed that if an individual algorithm was applied to the three types of events the accuracy of results could be affected.

Thus, we used three algorithms to compare the prediction results to better understand netizens' behaviors and attitudes regarding the three event types.

In our experiment, every 10 min is an observation time slot, marked with t (t is an integer from 1 to n). We defined the popularity of event i in the training set (from event 1 to M) at time series t as $\mathrm{x_i}(\mathrm{t}|\mathrm{t} \geqslant \mathcal{T})$. The popularity of event i is given as $\mathrm{x_i}(\mathrm{t}|\mathrm{t} < \mathcal{T})$. Furthermore, $x[i]$ is the popularity increment during the ith time interval length T_s $x[i] = x(i * T_s - x((i-1) * T_s)$.

4.1. Linear algorithm

A linear algorithm is an easy and effective method for prediction, and is commonly used in popularity predictions for news and online content (Atal, 2005; Lerman and Hogg, 2010; Szabo and Huberman, 2010).

Linear prediction has two main advantages. First, the variables used in an unknown event are easily determined from the given variables. Second, it is not necessary to make any assumption as to which of the individual characteristics would be effective. However, a limitation of the method is that the popularity of the present event is predicted based on the popularity of data in the training set at the same time interval. We cannot match the growth pattern while ignoring its time phase.

Linear prediction is based on a time series. The present event's popularity for the next time series is predicted as a linear combination of the popularity in the training dataset at the same time interval.

Time series popularity of every event in our training set will be given a corresponding weight as the coefficient to predict the present event G. The popularity of the predicted event G from time 1 to t is given. The number of follow-ups from time 1 to time t $(\mathrm{t}|\mathrm{t} \leqslant \mathcal{T})$ $\mathcal{T}$ of event G is used to predict the time t popularity $\mathrm{y_G}(\mathrm{t}|\mathrm{t} \geqslant \mathcal{T})$ as follows:

$$\begin{aligned}
&\min \sum_{n=1}^{T_c} \left(\sum_{i=1}^{M} x_i[t] \times \omega_i - y_G[t] \right)^2 \\
&s.t., \sum_{i=1}^{M} \omega_i = 1 \\
&\omega_i \geqslant 0, \quad i = 1, \ldots, M
\end{aligned}$$

Based on the popularity within time 1 to t of the training set events, we used the above formula to minimize the approximate error to obtain the prediction coefficients to predict the popularity of the present event $\widehat{y_G}$. Furthermore, $\widehat{y_G}$ is based on the weighted sum of the popularity of the training set (event 1 to M) at time t + 1 as follows:

$$\widehat{y_G}[T_G + j] = \sum_{i=1}^{M} x_i[T_G + j] \times \omega_i[T_G + j]$$

4.2. Correlation-based algorithm

In the linear algorithm, we compare $\mathrm{y_G}(\mathrm{t}|\mathrm{t} \geqslant \mathcal{T})$ with $\mathrm{x_i t}(\mathrm{t}|\mathrm{t} \geqslant \mathcal{T})$ to obtain the weight ω_i of each event. The popularity of $\widehat{y_G}$ for the time T interval is calculated from the weighted average of the popularity of training set events at time T.

However, a correlation-based algorithm searches all the popularity trends of all events and finds the part most closely correlated with the given part of the present event, and extrapolates the prediction value.

Correlation-based algorithm does not match the predicted event with the training set events at the same time interval. Instead, the algorithm searches the popularity of all events in our training set, and finds the part of event i that shows the closest similarity in shape with the given popularity of the present event $\mathrm{y_G}$. The popularity of event i for the following time $\mathrm{x_i(t)}$ is used to predict $\widehat{y_G}$.

The correlation function between the popularity of the present event and each event in the training set is calculated, as is the maximum value $\mathrm{v_i}$ and its index $\mathrm{k_i}$ as follows:

$$\begin{aligned}
&k_i = \arg \quad \max{}_1 \sum_{n=1}^{T_G} x_i[n + j] y_G[t] \\
&v_i = \sum_{n=1}^{T_G} x_i[n + k_i] y_G[t]
\end{aligned}$$

The popularity of the present event $\widehat{y_G}$ at time $\mathrm{T_G}$ is predicted as shown below:

$$\begin{aligned}
&\widehat{y_G}[T_G + d] = \frac{1}{M} \sum_{i=1}^{M} x_i[T_G + k_i + d] \omega_i \\
&\omega_i = \sum_{n=1}^{T_G} y_G[n] y_G[n] / v_i
\end{aligned}$$

4.3. Modified state transition-based algorithm

The linear algorithm's prediction matches the present event with all those in the training set with the same time series. The correlation based-algorithm searches each event's popularity in the training set in all time sequences. Neither method categorizes the online behavior over time to reveal distinct patterns of popularity growth.

Using a modified state transition-based algorithm, we categorized the behavior of online attention over time to reveal distinct patterns of popularity growth. We used the affinity propagation algorithm (AP) (Frey and Dueck, 2007) to cluster and handle non-linear dependencies between data and to reduce the complexity of the training data. AP algorithm considers all data points at the beginning stage and pass message information between data or nodes. AP algorithm calculates similar nodes, converges similar nodes, and discards ones that exhibit significant differences.

To decrease the computation complexity, sparse matrix representation is used to reduce nodes that exchange messages that reduce the message exchange while ignoring some data when similarity drops below a given level.

In modified clustering algorithm, a set of events consists of particular window $W = [\frac{T}{l}]$. The popularity of each time window is collected and AP algorithm is applied to find the node that can represent the set data in this window, or a transition from one state to another. The similarity between windows is calculated using the following formula:

$$\text{Similarity } (\omega_\mathrm{i}, \omega_\mathrm{j}) = \frac{\sum_{\mathrm{i=n}}^{l} \omega_\mathrm{i}[\mathrm{n}]\omega_\mathrm{j}[\mathrm{n}]}{\max\left\{\sum_{\mathrm{i=n}}^{l} \omega_\mathrm{i}[\mathrm{n}]\omega_\mathrm{i}[\mathrm{n}], \sum_{\mathrm{i=n}}^{l} \omega_\mathrm{j}[\mathrm{n}]\omega_\mathrm{j}[\mathrm{n}]\right\}} = -1$$

With this method, each set of events is organized based on their similarity of growth pattern or popularity development. The prediction is based on the similar group's growth pattern.

5. Results

5.1. Descriptive results

Table 1 displays the descriptive findings from the collective hot events data of Sina Weibo. Among the hot events, more than half are non-public events. This result actually is consistent with Loader and Mercea's statement. They point out that social media's dominant uses are entertainment, consumerism, and content sharing among friends (Loader and Mercea, 2011). But the average number of follow-ups for non-public events is far less than that of public events. One third of total events are social public events, which have the highest average number of follow-ups. Political hot events are least represented with just over 10% of the total, but the average number of follow-ups for political events is very close to that of social events. Concerning public events, the total number of public events, including political events and social public events, is 1897, accounting for 46% of the total. From the angle of public sphere, social media offers increasing opportunities for political and social communication and enable democratic capacities.

5.2. Prediction results

We used the three algorithms addressed in Section 2 to predict the popularity of three kinds of hot online events. Comparing the results (Fig. 1), we found that the prediction results of the modified state transition-based algorithm were the most accurate, followed by those of the correlation algorithm, and then the linear algorithm.

The prediction results for the three event types were similar using the linear algorithm. This was also true of the correlation algorithm. There is very little difference among the popularity predictions of the three event types (see Figs. 2 and 3). After 19 h, the prediction accuracy for the three types of events using the linear algorithm is approximately 25% (24% for non-public events, 25% for social events and 26% for political events). The accuracy of the correlation algorithm is 36% for all three types of events.

The linear algorithm predicted the current variable from the weighted average of the data in the existing set. The convergence of the three types of events with the linear algorithm shows little difference in prediction accuracy for the three types.

The correlation algorithm detects the popularity of every single event to find the one that is most similar with the present event. The prediction accuracy lines almost overlap and the overall results improve around 10%. Thus, matching the present event with a single event is very accurate.

The popularity prediction results of the three type events using the modified state transition-based algorithm show distinct separation (Fig. 4). From the beginning, the political events popularity prediction shows better accuracy than social events and non-public events. Regarding maximum values, the prediction accuracy of political events is 10% higher than that of social events and non-public events. After 19 h, the prediction accuracy remains at 34% for non-public events, 37% for

Table 1
Descriptive data for three types of online hot events.

	Political events	Social public events	Non-public events	All events
Number of events	513	1384	2238	4135
Average follow-up	2046	2171	1656	1877*
Percentage of total events	12.4%	33.5%	54.1%	100%

* The average follow-up of all events is the weighted average of the three types of events.

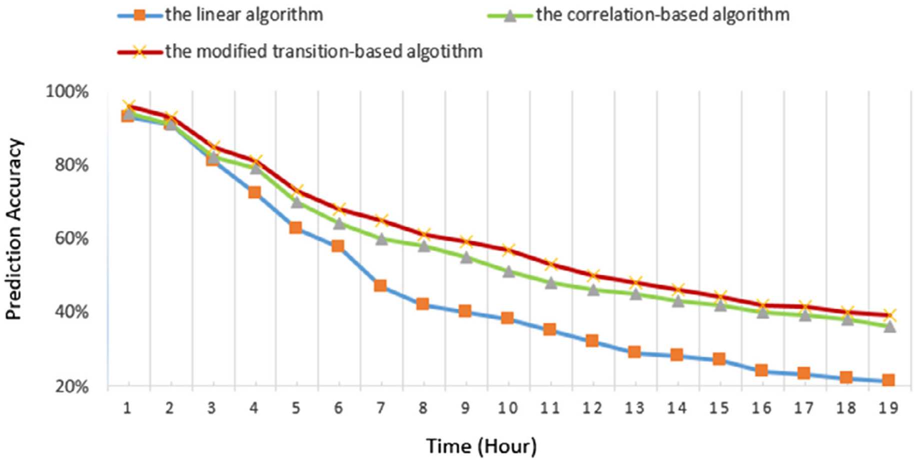

Fig. 1. Average prediction results with different algorithms.

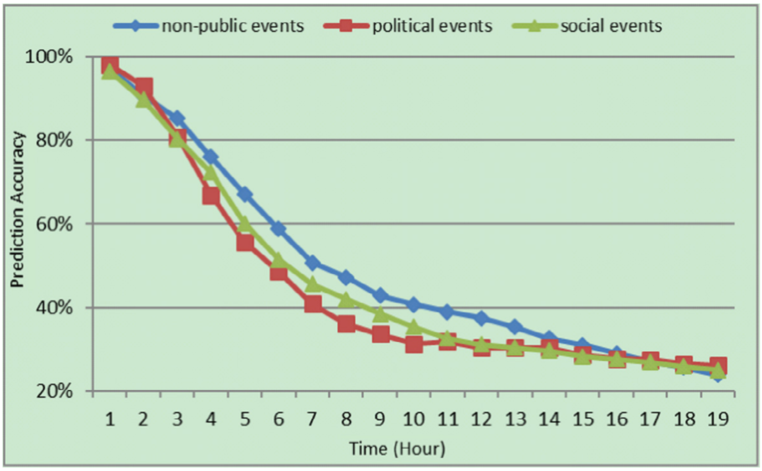

Fig. 2. Popularity prediction with the linear algorithm.

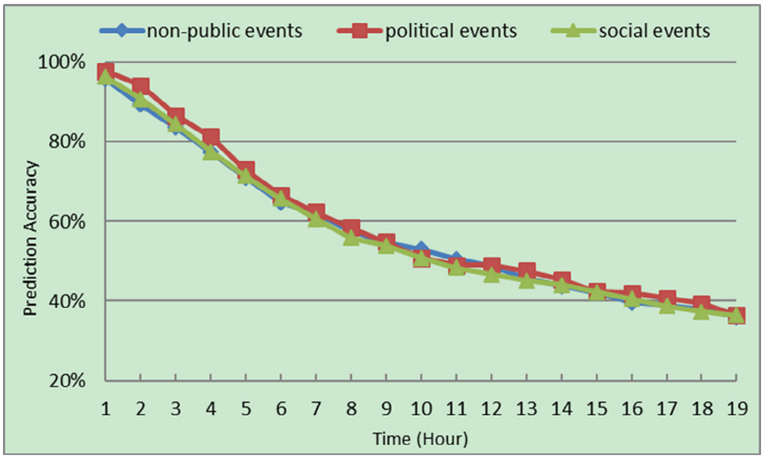

Fig. 3. Popularity prediction with the correlation-based algorithm.

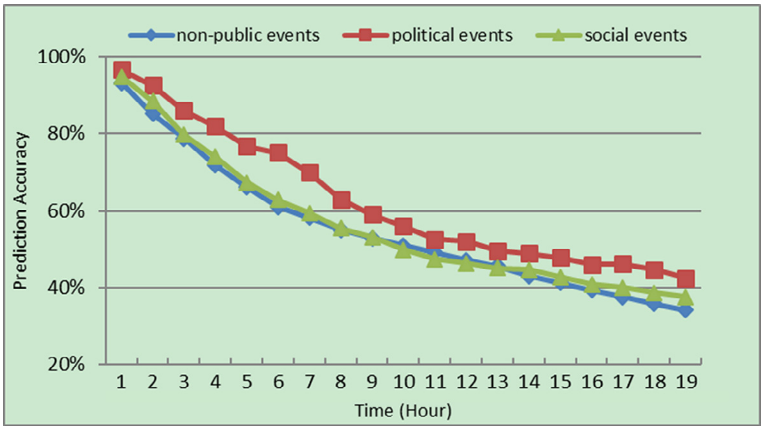

Fig. 4. Popularity prediction with the modified transition-based algorithm.

social public events and 43% for political events. Compared with the correlation algorithm, the accuracy of prediction for non-public events slightly decreases, social events shows little change, but that for political events increases by 7%.

6. Discussion and conclusion

We conduct an interdisciplinary study on online collective behaviors based on Sina Weibo, the largest microblogging provider in China. A Microblog is a platform for users to express their own opinions about any topic imaginable. Social media's dominant uses are entertainment, consumerism, and content sharing among friends (Loader and Mercea, 2011). But such topics are unlikely to attract the interest of others and become hot events. Thus, the comparable largest portion of hot events among each category does not truly show people's preferences. We noticed that non-public events have on average less follow-up threads than public events. The number of follow-ups to some extent shows people's concerns and is a good indicator of their preferences. Compared with non-public events, for example, anecdotes, box news and business and commercial news, internet users show an obvious desire to participate in public events. In this research, the average number of follow-ups regarding political events is similar to that for social events in China. This perhaps indicates that Chinese pay a similar level of attention to political and social topics. From the angle of public sphere, social media offers Chinese the increasing opportunities for political and social communication and enable democratic capacities.

In an authoritarian country, at least, this partly shows that people have a desire to participate in public affairs, once they have access to the means to express themselves. The result is consistent with previous empirical offline studies that have verified that the Chinese support democratic values and freedom (Chen and Zhong, 1998). A recent survey conducted in five of China's biggest cities also found that Chinese are concerned about public affairs, but more so regarding affairs closely related with day-to-day living such as housing price, inflation, medical care, education, income, and social welfare. Contrary to the high interest in social affairs, people show less interest in political reform, human rights, and freedom (Zhong and Chen, 2013).

China has historically been regarded as an interesting and complicated case of collectivism (Triandis, 1995). Traditional Chinese culture may be inclined toward verticality, but the State advocates horizontal themes. China may be considered a horizontal collective culture, influenced by the historical political and economic environment that emphasized egalitarian and group centered values. Social events are concerned with people's personal interests and benefits, so it is understandable that they draw greater public attention. As far as social affairs are concerned, this result may be consistent Sivadas' measurement, China is vertical individual country (Sivadas et al., 2008) Our survey found that people also maintain a similar interest in political events. One possible reason is that some political issues, such as corruption, also relate to people's individual interests. Bureaucratic corruption violates public benefits, so concern about corruption reflects people's concerns about their own interests. Another reason is that online media is a freer public space, people could express themselves more freely online than offline. For participation in offline events, people are more passive; they express their ideas only when their interests are harmed. Regarding political events, they are more sensitive and cautious about expressing ideas and taking action. However, online events allow people to easily follow others' comments and actions. Thus, people show a similar interest in political and social events.

The linear algorithm offers predictions based on the weighted average popularity of the training dataset. The weights are obtained by calculating the minimum square error of the given popularity of the present event and the events in the dataset. It includes all the data in the set and does not differentiate people's behavior. A correlation algorithm compares the given popularity of the present event with every event's popularity in the dataset, but it does not compare them at the same time interval. Instead, it matches the most similar part and uses the given popularity to predict the present event. In contrast, the modified state transition-based algorithm clusters the netizens' behavior using affinity propagation, and then matches the present event with the cluster that is the most similar. The prediction result with this algorithm for non-public events and social events shows little change, while it significantly improves the prediction results for political events. Here, the biggest difference for state transition-based algorithm is that it clusters the data based on similarity. The prediction is established on the similarity between the present event and the clusters.

The modified state transition-based algorithm considers clusters of behavior (the herd effect) and categorizes events based on likeness. This algorithm matches the current event with a group of events to which it is most similar, and uses the popularity of this similar group to predict the popularity of the current event. The more similar the events' popularity growth, the easier it is to cluster them together. If the present event finds a similar popularity growth cluster, the prediction result is more accurate. Thus, with this algorithm, people's behavioral convergence may lead to accurate prediction results.

The improved result for political events using the modified state transition-based algorithm shows that people's behavior regarding political events is more clustered and convergent. Although social public events are the most common, such collective events pose no threat to fundamental societal stability. Yu (2000) argues that the nature of social conflict in China is interest conflict. He explains the number of collective events are increasing annually, but 80% of collective events concern social interest conflicts (Yu, 2008). They concern the safeguarding of participants' rights. In other words, such events derive from conflicts between people's interests and government, rather than conflicts of fundamental values or beliefs. Participants merely want to safeguard their own rights and interests according to the current legal system and have no intention to challenge current legal and social rules. Collective behaviors are more passive than active. People only respond when their interests are harmed and their needs are not satisfied. They do not take such action for political purposes. Thus, current collective events concern social issues not political ones. It is for this reason that Chinese society remains stable while facing mounting social problems. Although the results of our study support this point, our findings also show that people's clustering behavior is becoming more convergent on political events. Thus, more attention should be paid to this particular characteristic. Although as far as the number of follow-up is concerned, we observed vertical individualism intention for online social and political behavior. Based on the better results for political events using the modified state transition-based algorithm, people show collectivism on political affairs than social affairs.

We now conclude the paper. Our research contributes to the online events prediction in the following four folds. First, our social-computational study on popularity predictions of social online hot events offers predictions for online events. Therefore, tracking the development modes of online hot events and predicting their growth patterns can help us detect people's behavior offline.

Second, for an in-depth study on China netizens' concern with public affairs, we divided the events into three categories. Such classification helps us better observe Chinese attitudes and behaviors regarding public issues. Some earlier studies investigated Chinese participation in public affairs, but focused primarily on offline behavior and some collective events. There have also been some investigations into online electoral prediction, but scant attention has been paid to online public participation and popularity predictions for online public affairs. The present study is the first to examine people's online behavior and predict the popularity of public events using online data from China. As an authoritarian country, the level of Chinese public participation is not as high as in some democratic countries. Understanding people's attitudes and behavior toward political affairs and social affairs online is essential to understand Chinese society.

Third, unlike other studies dealing with public participation, our research is an interdisciplinary and empirical study based on a unique large dataset from China's largest and most popular blogger provider (Sina Weibo). We used topic-ranking lists from different search engines to define hot events, and analyzed over 14 million posts on these topics on Sina Weibo. Using such a large dataset ensures that our data covers almost all hot events in that period, and makes our research more relevant.

Fourth, three algorithms (linear, correlation-based and modified state transition-based) are used for prediction in our study. Compared with online survey methods, which require careful consideration of conceptual and methodological issues (Jansen et al., 2007), our research collected netizen's online behavior data directly and used three different algorithms to analyze and predict online behavior. The results compare the prediction accuracy of the three algorithms in predicting the popularity of the three types of events. The mathematical aspects of the three algorithms can be used to explain and observe people's behavior to some extent. Our computational social science study analyzed a large-scale dataset of over 14 million hot online threads relating to hot online events from the viewpoint of follow-up statistics and algorithmic predictability. We found that non-public events represent the largest proportion of hot events in terms of the number of topics, but do not receive the greatest level of attention. The highest average for follow-up threads concern social events, showing that people in China place their personal interests and concerns first and the society is inclined to be vertical individualism in social affairs. Thus, the key reason behind collective events is conflicts regarding individual interests rather than those concerning values or beliefs. Chinese is more vertical collectivism for political affairs. Although major social conflicts are not considered value conflicts, such concern has its own clear mode. There may be a special group of people who pay more attention to political events and take similar action. To some extent, understanding their behavior model will help us to better predict the popularity of online hot political events.

Due to the limited resources, we only classified all events into three broad categories, and did not further divide them in subcategories. We believe more detailed classification can help us understand Chinese online collective behavior better. Another influence about our research is from the strict online contents censorship in China. The purpose of the censorship program is to reduce the probability of collective action. So our research could not reveal a full picture of people's online behavior. In our future work, we will collect even bigger and more diverse data related to online hot events and applied even more advanced prediction methods to analyze online collective action.

References

Applegate, D., Archer, A., Gopalakrishnan, V., Lee, S., Ramakrishnan, K.K., 2010. Optimal content placement for a large-scale VoD system. In: Paper Presented at the Proceedings of the 6th International Conference.

Atal, B.S., 2005. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. 55 (6), 1304–1312.

Barme, G.R., Davies, G., 2005. Intellectual contestation and the Chinese web. Chin. Intellect. Between State Market 75.

Bennett, W.L., Iyengar, S., 2008. A new era of minimal effects? The changing foundations of political communication. J. Commun. 58 (4), 707–731.

Chen, J., Zhong, Y., 1998. Defining the political system of post-Deng China: emerging public support for a democratic political system. Problems Post-Communism 45 (1), 30–42.

Chiou, J.-S., 2001. Horizontal and vertical individualism and collectivism among college students in the United States, Taiwan, and Argentina. J. Social Psychol. 141 (5), 667–678.

CNNIC, 2015. Statistical Report on Internet Development in China. China Internet Network Information Center, Beijing.

Cui, E., Tao, R., Warner, T.J., Yang, D.L., 2014. How do land takings affect political trust in rural China? Political Studies.

Duan, W., Gu, B., Whinston, A.B., 2005. Analysis of herding on the internet—an empirical investigation of online software download. Paper presented at the AMCIS.

Dumais, S.T., 2004. Latent semantic analysis. Annu. Rev. Inf. Sci. Technol. 38 (1), 188–230.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. Science 315 (5814), 972–976.

Habermas, J., 1991. The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society. MIT Press.

Habermas, J., 2006. Political communication in media society: Does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research. Commun. Theory 16 (4), 411–426.

Ho, P., 2014. The 'credibility thesis' and its application to property rights:(in) secure land tenure, conflict and social welfare in China. Land Use Policy 40, 13–27.

Hsu, C.L., Chuan-Chuan Lin, J., Chiang, H.S., 2013. The effects of blogger recommendations on customers' online shopping intentions. Internet Res. 23 (1), 69–88. http://dx.doi.org/10.1108/10662241311295782.

Huang, P.C., 1993. "Public Sphere"/"Civil Society" in China?: the third Realm between state and society. Modern China, 216–240.

Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A., 2009. Twitter power: Tweets as electronic word of mouth. J. Am. Soc. Inf. Sci. Technol. 60 (11), 2169–2188. http://dx.doi.org/10.1002/asi.21149.

Jansen, K.J., Corley, K.G., Jansen, B.J., 2007. E-survey methodology. Handb. Res. Electron. Surv. Measur., 416–425

Jarren, O., Donges, P., 2002. Politische Kommunikation in der Mediengesellschaft. Eine Einführung 1.

King, G., Pan, J., Roberts, M., 2012. How censorship in China allows government criticism but silences collective expression. Paper presented at the APSA 2012 Annual Meeting Paper.

Knight, J., 2013. The economic causes and consequences of social instability in China. China Econ. Rev. 25, 17–26.

Lee, E., Lee, B., 2012. Herding behavior in online P2P lending: an empirical investigation. Electron. Commerc. Res. Appl. 11 (5), 495–503.

Lerman, K., Hogg, T., 2010. Using a model of social dynamics to predict popularity of news. Paper presented at the Proceedings of the 19th International Conference on World Wide Web.

Lewis, J.W., Litai, X., 2003. Social change and political reform in China: meeting the challenge of success. China Quart. 176, 926–942.

Liang, A.R.-D., 2014. Enthusiastically consuming organic food. Internet Res. 24 (5), 587–607. http://dx.doi.org/10.1108/IntR-03-2013-0050.

Liu, Z., Jansen, B.J., 2014. Predicting potential responders in social Q&A based on non-QA features. Paper presented at the CHI '14 Extended Abstracts on Human Factors in Computing Systems, Toronto, Ontario, Canada.

Loader, B.D., Mercea, D., 2011. Networking democracy? Social media innovations and participatory politics. Inf. Commun. Soc. 14 (6), 757–769.

Muller, E.N., Jukam, T.O., 1977. On the meaning of political support. Am. Polit. Sci. Rev., 1561–1595

Onnela, J.-P., Reed-Tsochas, F., 2010. Spontaneous emergence of social influence in online systems. Proc. Natl. Acad. Sci. 107 (43), 18375–18380.

Qiu, L., Lin, H., Chiu, C.-Y., Liu, P., 2014. Online collective behaviors in China: dimensions and motivations. Anal. Social Issues Public Policy. http://dx.doi.org/10.1111/asap.12049.

Salton, G., McGill, M.J., 1983. Introduction to modern Information Retrieval.

Shirk, S.L., 2007. China: The Fragile Superpower. Oxford University Press.

Sivadas, E., Bruvold, N.T., Nelson, M.R., 2008. A reduced version of the horizontal and vertical individualism and collectivism scale: a four-country assessment. J. Bus. Res. 61 (3), 201–210.

Sullivan, J., 2014. China's Weibo: is faster different? New Media Soc. 16 (1), 24–37.

Szabo, G., Huberman, B.A., 2010. Predicting the popularity of online content. Commun. ACM 53 (8), 80–88.

Triandis, H.C., 1995. Individualism & Collectivism. Westview Press.

Triandis, H.C., 2004. The many dimensions of culture. Acad. Manage. Execut. 18 (1), 88–93.

Triandis, H.C., Gelfand, M.J., 1998. Converging measurement of horizontal and vertical individualism and collectivism. J. Pers. Soc. Psychol. 74 (1), 118.

Wang, Y., 2001. Research on fundamental problems of public administration. China Admin. Manage. 11, 1.

Yang, G., 2013. The Power of the Internet in China: Citizen Activism Online. Columbia University Press.

Yu, J., 2000. Interest, authority and order: analysis on collective events of villagers against local government. China Rural Surv. 4, 70–76.

Yu, J., 2008. Dilemma of controlling social revenge events in China. Contemp. World Socialism (1), 4–9.

Zheng, Y., Wu, G., 2005. Information technology, public space, and collective action in China. Comparative Political Studies 38 (5), 507–536.

Zhong, Y., Chen, Y., 2013. Regime support in urban China. Asian Survey 53 (2), 369–392.