

# Topic dynamics in Weibo: a comprehensive study

Rui Fan<sup>1</sup> · Jichang Zhao<sup>2</sup> · Ke Xu<sup>1</sup>

Received: 21 December 2014 / Revised: 29 June 2015 / Accepted: 1 July 2015 / Published online: 14 July 2015  
© Springer-Verlag Wien 2015

**Abstract** The tremendous development of online social media has changed people's life fundamentally in recent years. Weibo, a Twitter-like service in China, has attracted more than 500 million users in less than 5 years and produces more than 100 million Chinese tweets everyday. In these massive tweets, different user interests and daily trends are reflected by different topics. To our best knowledge, a systematic investigation of topic dynamics in Weibo is still missing. Aiming at filling this vital gap, we try to comprehensively disclose the topic dynamics from the perspective of time, geography, demographics, emotion, retweeting and correlation. An incremental learning framework is first established to probe more than 200 million streaming tweets and an interaction network constituted by around 90,000 active users. Many interesting patterns are then revealed, which could provide insights for topic-related applications in online social media, such as user profiling, event detection, trend tracking or content recommendation.

**Keywords** Topic classification · Topic dynamics · Topic patterns · Topic correlation · Weibo

## 1 Introduction

In these days, the online social media attract a large number of users and produce abundant content everyday. Twitter, launched in 2006, is a microblogging service and severely restricts its textual content named tweet to less than 140 words. In China, Weibo, a service like Twitter, deployed in 2009, has more than 500 million users in less than 5 years and its massive users post more than 100 million Chinese tweets everyday. These tweets indeed provide a “big data” window to the study of detailed human behavior from an unprecedented perspective. Specifically, tweets in Twitter-like services real-timely reflect the real-world circumstance and deliver complicated signals of the present; for example, when a hot social event emerges, users will post, retweet and comment about the event and produce lots of tweets in a short period. These accumulating tweets would then form a new trend of the entire Internet instantly, which can be effectively reflected by different topics and their evolution.

Evan Williams, one of co-founders of Twitter, defines the service as “an information network,” and Twitter itself also changes the question that it asks users from “What are you doing?” to “What's happening?”. Considering its worldwide distributed network of users of any age and social backgrounds, it represents a “low-level news flash portal” (Cataldi et al. 2010). Nowadays, Twitter is becoming a big library of up-to-date topics about real-world events (Kwak et al. 2010; Song et al. 2012). Similarly, Weibo is growing to be the most pervasive channel of message exchange in China. From this perspective, Twitter or Weibo is a huge information library which contains abundant user-generated news of various topics.

In the era of online social media, information overload is becoming a more and more serious problem. Everyday we

---

✉ Jichang Zhao  
jichang@buaa.edu.cn

<sup>1</sup> State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, People's Republic of China

<sup>2</sup> School of Economics and Management, Beihang University, Beijing 100191, People's Republic of China

receive many redundant messages or news feed that we are not interested in. As a classical solution, text classification is employed to overcome this trouble by dividing content into different topics (Joachims 1999). Several interesting applications were also built toward this issue; for example, Jagan et al. developed *TwitterStand* from Twitter tweets (Sankaranarayanan et al. 2009), which is a news-processing system that picks news-related tweets from the disordered corpus. Ritter et al. extract events from Twitter and assign a label for each of them, such as *Sports* or *Politics*<sup>1</sup> (Ritter et al. 2012). Weibo itself also build a system to classify some influential tweets into several categories, such as sports, movies and entertainment. All of these works have made great effort to classify tweets or extract certain types of information. However, mining topics from online social media is a significant but challenging task; for example, the limit of 140 characters brings the issue of extreme sparsity in text. Moreover, people usually use abbreviations, neologism, external URLs or pictures to express complementary information, which lead to the data disorder. Particular, little work has been done to systematically investigate the dynamics of different topics from a comprehensive view. For Twitter, Yang et al. only studied the temporal pattern of the online content and how the content's popularity grows and fades over time (Yang and Leskovec 2011). Regarding Weibo, little attention has been paid by the research community, especially from the view of sentiment and geography. Meanwhile, the difference of topic dynamics between different cultures is not revealed clearly until now.

In order to fill the above void, we build a fast and convincing topic classifier on Weibo to understand the topic dynamics of online social media. To our best knowledge, this is the first work to explore the temporal pattern, spatial pattern, gender pattern, sentimental pattern, subtopic pattern, mobility pattern, retweeting pattern and correlation pattern of Weibo topics. These patterns are revealed from two datasets, which include more than 200 million streaming tweets posted in the year of 2013, and an interactive network of users and their tweets posted in 2013. Our main contributions are listed as follows:

- A complete taxonomy for topics in Weibo is established, including *Society*, *International*, *Sports*, *Technology*, *Entertainment*, *Finance* and *Military*. Then, an efficient topic classifier with strategy of incremental learning is built, which achieves satisfactory accuracy and *F*-measure (both around 84 %).
- More than 200 million tweets from January 6, 2013 to November 30, 2013 in Weibo are studied comprehensively. Many interesting patterns in topic dynamics are

revealed; for instance, *Entertainment* occupies the largest fraction in Weibo and the sentiment distribution in different topics varies greatly. And the analysis of spatial pattern shows that the number of *Entertainment* tweets in Hong Kong and Taiwan far exceeds that of other provinces. Tweets of most topics except *Entertainment* decline on weekends, and *Finance* shows an obvious periodic pattern and surprisingly strong interactive correlation.

- A Weibo interaction network constituted by around 90,000 active users and their 100 million tweets is thoroughly explored. Many novel patterns are disclosed from views of subtopics, social mobility, retweeting and correlation. We find that some hot subtopics discussed across several topics, *Sports* messages always propagate through close friends, and *Entertainment* and *Society* are more easily be interested by the same user.

## 2 Related works

Existing text categorization studies in Twitter have different definitions of tweets' topics. In Sriram et al. (2010), the authors classify tweets into News, Events, Opinions, Deals and Private Messages. And they classify tweets into these categories by using authors' information and tweets' features. Becker et al. divide tweets to messages about real-world events and nonevent messages based on analyzing the stream of Twitter messages (Becker et al. 2011). Sankaranarayanan et al. build a system called TweetStand to classify tweets as news and non-news (Sankaranarayanan et al. 2009). Meanwhile, there are some other works that utilize external knowledge to handle the sparsity of short tweets. Wikipedia knowledge is widely used as the external information (Schönhofen 2009; Banerjee et al. 2007). Specifically, Genc et al. compute the most similar Wikipedia page of each tweet and employ this external information to cluster tweets into some different categories (Genc et al. 2011). External hyperlinks are also employed to obtain the topic of individual tweet in Kinsella et al. (2011). However, obvious drawbacks still exist in the previous literature; for instance, few of them define a complete taxonomy, some unreliable information might be brought into classification from the exterior, and a simple and feasible solution for massive tweets is still missing.

As for topic dynamics, Quercia et al. studied topics, geography and emotion in Twitter from the view of *network metrics*. They want to testify whether existing sociological theories of real-life networks hold in Twitter (Quercia et al. 2012). Some researchers also analyzed the temporal aspect of trends and trendsetters in Weibo and compared it with the earlier observations from Twitter (Yu

<sup>1</sup> <http://statuscalendar.com>.

et al. 2013). To investigate the popularity of Twitter “trending topics,” Sebastien et al. provided a rigorous temporal and spatial analysis (Ardon et al. 2013). However, those previous works do not focus on topics based on media feature of Twitter-like services and a more comprehensive mining of topic dynamics from perspectives such as gender, emotion and interaction still deserves further explorations.

Topics of specified users also attract attention in the existing studies. Yamaguchi et al. propose an approach to discover appropriate topics for a user by using his/her Twitter list, which is an official functionality to represent topics that users are interested in Yamaguchi et al. (2011). Michelson and Macskassy make efforts on discovering Twitter users’ topics of interest by examining the entities they mention in their own tweets (Michelson and Macskassy 2010). Bogdanov et al. propose a model called *Genotype* which can generate a “per-topic summary” of a user’s interest (Bogdanov et al. 2013). Actually, interest preference can also be demonstrated through the classification of Twitter users; for example, Yang proposed a method that divides Twitter users into different clusters based on the periodicity, which inspires the present work (Yang et al. 2013).

Since retweeting is the key mechanism for information diffusion in Twitter (Suh et al. 2010), it is intuitively related to topic evolution. Boyd et al. examine the behavior of retweeting as a way by which participants can be “in a conversation” (Boyd et al. 2010). Yang et al. build a factor graph model to predict Twitter users’ retweeting behaviors (Yang et al. 2010). Little attention has been paid to how it affects the patterns in topic dynamics. In this paper, we study this issue by constructing an interaction network with retweeting messages.

### 3 Topic classifier

In this section, we build a convincing and fast classifier to divide millions of tweets from Weibo into seven topics automatically. Particularly, our learning framework supports the mechanism of incremental training, which could keep the classifier’s performance stable.

#### 3.1 Crowdsourcing-based labeling

As the basis of all, a topic classifier with high reliability and satisfactory efficiency is a primary start. Considering the news feature of Weibo, we browse many news media such as “Google News,” “NetEase News” and “Sina News.” Combining with the analysis of our Weibo data, we propose our taxonomy of topics including *Society*,

*International*, *Sports*, *Technology*, *Entertainment*, *Finance* and *Military*. We argue that this is a complete classification and that it can cover the majority of tweets. In fact, such taxonomy reflects the media property of Weibo to some extent. After determining the taxonomy, we try to gather the labeled data as the training data. It is obvious that manually annotating tweets is not feasible because of the massive dataset and its extreme disorder. So we adopt a crowdsourcing-based method. We notice that there are some professional accounts that only post a certain kind of tweets. To be specific, the user named “Sina Sports” only posts messages that are related to the sports information. Moreover, these accounts are hosted by professional reporters. Therefore, the tweets labeled (posted) by them are convinced to belong to a certain category. Based on this assumption, we pick out several authoritative users for each category and then crawl all their tweets. Using this method, we collected an available training set which includes 411,373 tweets. Table 1 exposes more details about this labeled data. Meanwhile, comparing with the ordinary tweets, messages in this corpus are also short texts that obey the limitation of 140 Chinese characters, which further guarantees the uniformity of the training samples. Furthermore, these specific accounts will produce new tweets everyday, which means we can easily train an incremental classifier to capture the newly emerging features. We have made our dataset publicly available, and it can be downloaded freely through <http://goo.gl/Q1oeNX>.

#### 3.2 Classification framework

To find the trends of Weibo from the view of topics, a certain category should be assigned to each tweet. Based on the labeled data, we build a framework to achieve this goal. Assume that  $C = (c_1, c_2, \dots, c_7)$  is the set of topics,  $c_i$  is an arbitrary topic,  $d$  represents a tweet, and it can be split as  $(w_1, w_2, \dots, w_n)$ , where  $w_i$  is a term. In this paper, Bayesian classifier, a classical probability method, is used to accomplish the classification task. The formula to determine the label of a tweet is

$$c = \text{*arg max}_{c_j \in C} P(c_j|d), \quad (1)$$

then, the equation is rewritten as

$$c = \text{arg max}_{c_j \in C} \frac{P(w_1, w_2, \dots, w_n|c_j)P(c_j)}{P(w_1, w_2, \dots, w_n)}. \quad (2)$$

For all categories, the expression  $P(w_1, w_2, \dots, w_n)$  is the same.  $P(c_j)$  is the prior probability of one class, and we set it the same value for all categories. So we only need to calculate  $P(w_1, w_2, \dots, w_n|c_j)$ . Based on the independence assumption, this equation is equal to  $\prod_{i=1}^n P(w_i|c_j)$ , in which

**Table 1** The number of labeled tweets and the precision, recall and  $F$ -measure of the cross-validation

Topic	Counts	Precision (%)	Recall (%)	$F$ -measure (%)
Society	22,975	65.31	74.71	69.69
Finance	66,134	87.04	86.77	86.90
Military	34,617	90.04	92.43	91.22
Entertainment	91,679	88.53	95.33	91.80
International	14,253	65.83	59.00	62.23
Sports	108,041	98.62	93.90	96.20
Technology	73,674	92.36	86.83	89.51
All	411,373	83.96	84.14	84.05

$$P(w_i|c_j) = \frac{n^{c_j}(w_i) + 1}{\sum_q (n^{c_j}(w_q) + 1)}, \quad (3)$$

where  $n^{c_j}(w_i)$  is the times that the word  $w_i$  occurs in all tweets of  $c_j$ . We use *Laplace smoothing* to avoid zero probability.

### 3.3 Feature selection

Considering the high dimension of features (around 300,000 terms), we compare three feature selection methods depicted in Yang and Pedersen (1997) and Novakovic (2010) to reduce the features of our classifier and enhance the classifier's performance. Specifically, document frequency (DF) calculates the document frequency for each unique term in the training set and then removes terms whose document frequency is lower than a predefined threshold. Information gain (IG) measures the number of bits of information obtained for category prediction by calculating the presence of absence of a term in a document. CHI ( $\chi^2$ ) reflects the lack of independence between term  $t$  and category  $c$  and then compared to the  $\chi^2$  distribution with one degree of freedom. Its formula can be written as

$$\text{CHI} = \frac{N(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)}, \quad (4)$$

where  $A$  is the number of times  $t$  and  $c$  co-occur,  $B$  is the number of times the  $t$  occurs without  $c$ ,  $C$  is the number of time the  $c$  occurs without  $t$ ,  $D$  is the number of times neither  $c$  nor  $t$  occurs, and  $N$  is the total number of documents.

We randomly split our labeled dataset into training and test sets. The training data contain 90 % of the raw data. The result is shown in Fig. 1a. We can draw a conclusion that even when we just pick 20 % features, the  $F$ -measure still reaches 83 %. But different from Yang's work Yang and Pedersen (1997), the  $F$ -measure always raises when the number of features grows, while the point that it begins to

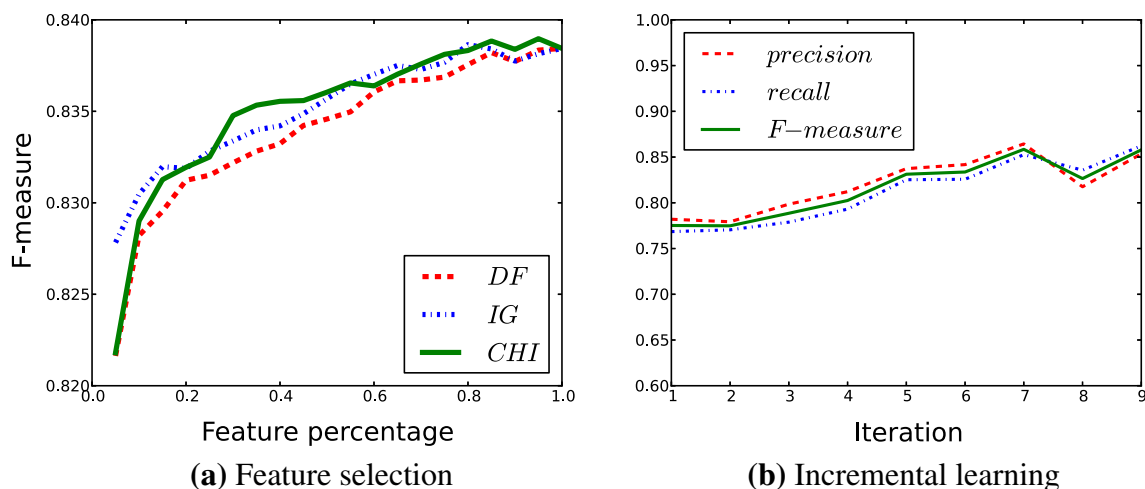
saturate or drop does not appear. The reason is that Weibo dataset consist of short- text documents, so features which are benefit to this classification task are always insufficient. So reserving a large amount of features is necessary, and therefore, we employ the full-feature set in the following experiments.

### 3.4 Cross-validation

In this section, we conduct the cross-validation to testify the performance of our classifier. Evaluation measures employed here are precision, recall and  $F$ -measure. We divide our labeled dataset that was crawled from the professional users into ten equal partitions, and in every iteration we use a random piece as the testing data, while the others are unified as the training data. After averaging the three measurements, the result is shown in Table 1. The averaged  $F$ -measure is 84.05 %, which indicates that our classifier has high performance, especially as considering the 140 characters limit of each tweet.

### 3.5 Incremental training

As discussed before, the tweets in Weibo are evolving simultaneously with the real-world events. That is to say, the writing style is not fixed and new terms or hot words would emerge continuously. In order to tackle the challenge brought by new features, we embrace the incremental training into our classifier to keep its features up-to-date. In the experiment, we first divide the labeled data to ten partitions consecutively according to the time sequence. Then, the first  $i$  pieces are used to train the classifier and the  $i + 1$  piece will be used to test it. As  $i$  grows, the classifier would be updated iteratively. As shown in Fig. 1b, the precision, recall and  $F$ -measure raise gradually with iteration  $i$ , which means our incremental training is effective. Regarding the application scenario, we continuously collect the labeled tweets from the professional accounts and use them to update the classifier.



**Fig. 1** The result of feature selections and the incremental learning

### 3.6 Classifier comparison

In the previous literature, many experiments were performed on the comparison of text classification algorithms (Dumais et al. 1998; Zhang and Oles 2001; Yang and Liu 1999). And their results all show that SVM outperforms other classification models. Based on the same dataset, we also compare our Bayesian classifier with SVM. The  $F$ -measure of SVM is 85.8 % and the Bayes classifier is 84 %, which means that the  $F$ -measure of SVM is trivially higher than that of Bayesian method. With respect to the learning cost, SVM consumes much more time. Moreover, it is not easy to embed incremental learning feature into SVM when new labeled data are achieved. Hence, given the excellent performance and ability of incremental training, Bayesian classifier is the appropriate choice for us in this paper.

## 4 Topic dynamics

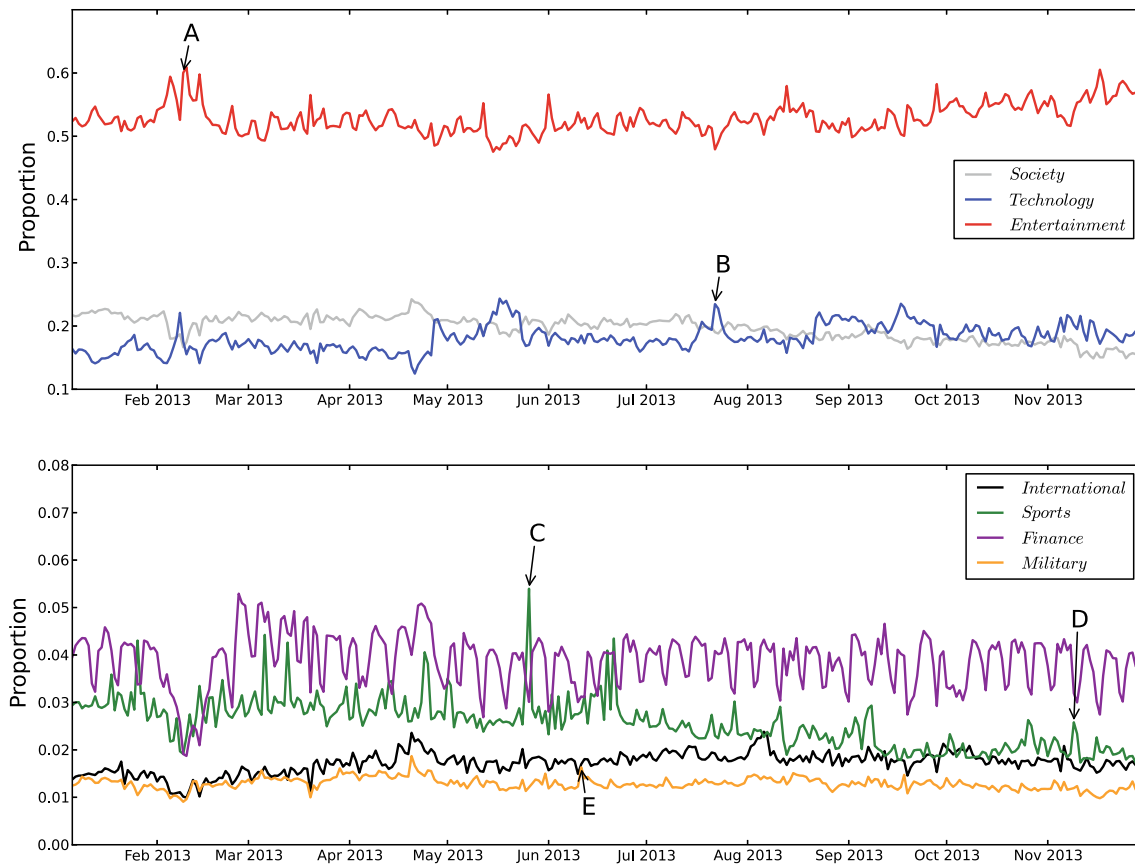
In this section, we investigate the topic dynamics in detail based on the topic classifier. We investigate the dynamics through disclosing the evolving patterns from the perspective of time, geography, gender, emotion, subtopic, retweeting and correlation. These studies are conducted on two datasets. The first one is a streaming data from January 6, 2013 to November 30, 2013 in Weibo and the total number of tweets is 238,517,038. These data are used for the study of temporal pattern, spatial pattern, gender pattern and sentimental pattern. The second one is an interaction network composed of 89,023 users and all of their tweets from 2009 to 2013 with the number of 123,727,746. Given its relative long period and detailed users' information, it can be used to discuss topic

dynamics from the view of long-term and social network. Hence, patterns of subtopics, social mobility, retweeting and topic correlation are systematically probed through this dataset. We obtain these data from Weibo through open APIs. The first dataset is crawled by the streaming API, which returns a sample of the most recent tweets. As a result, we can obtain a sample of the whole Weibo data, while the second dataset is crawled in two steps. First, a typical breadth-first search strategy is used to obtain the user set as in Fan et al. (2014) and Mislove et al. (2007). To be specific, we started our crawling task from several influential seeds (users verified by Weibo) and added their followers to our user set for future extension. This procedure was repeated and at last, we obtained 89,023 active users. Then, in the second step, we crawled their tweets published during a period of time. All tweets returned by the APIs contain abundant fields such as text, retweeting information and user information. Gender, location and followers' count are included in the user fields. As a result, we can use text to assign a topic for tweets, the gender and location information to study topic dynamics, and the retweeting message to build interactive network.

### 4.1 Temporal pattern

We discuss the temporal pattern of Weibo topics in this section. In the long period of time, we calculate the topics' daily proportions. The results are shown in Fig. 2. We find that *Entertainment* is the absolute majority among the seven topics and that it dominates Weibo, which is consistent with our intuition. Since after one day's hard work, Weibo users may always browse and retweet the entertainment tweets to relax themselves, and this behavior then stimulates the growth of entertainment tweets greatly.





**Fig. 2** The daily proportions of the seven topics and significant events are denoted by *different capitalized letters*

The phenomenon that the *Entertainment* messages occupy a surprisingly large proportion stimulates us to investigate whether this finding holds in Twitter, the majority of whose users comes from a different cultural background. Four hundred and sixty-seven million tweets posted by 20 million users covering a seven-month period from June 1, 2009 to December 31, 2009 are employed to perform the comparison. As estimated by the authors of Yang and Leskovec (2011), this is about 20–30 % of all public tweets published on Twitter during the particular time frame. Six categories of hashtags from Romero et al. (2011) are employed to determine the topic of a tweet from Twitter, including *Movie/TV*, *Technique*, *Political*, *Sports*, *Game* and *Music*. For a tweet containing hashtags belonging to just one topic, it would be labeled as this topic. The statistical result is listed in Table 2, from which we find *Entertainment* (*Movie/TV*, *Game* and *Music*) occupies a much smaller proportion than that in Weibo, just around 28 %, while out of expectation, *Political* tweets take the fraction of 29 %, which is significantly higher than that in Weibo (*Political* is not an explicit category in Weibo, while this type of tweets are impliedly included in *Finance*, *Military* and *International*). Thus, we assume that the fraction of political tweets in Weibo is

**Table 2** Proportions of topics in Twitter

Category	Number	Proportion (%)
Movie/TV	481,671	6.88
Technique	2,547,766	36.41
Political	2,037,654	29.12
Sports	389,522	5.57
Game	773,479	11.05
Music	767,777	10.97

roughly the sum over the above three categories). This result is consistent with the previous findings in Yu et al. (2013, 2011), and it also demonstrates the cultural difference that Chinese people pay more attention on entertainment affairs than West foreigners.

We can also learn from Fig. 2 that the proportion of each topic is changing over time. Intuitively, topics that Weibo users are talking are related to real-world events. With this in mind, we analyze the abnormal points in trends and try to mine the corresponding events. In early February (A), the proportion of *Entertainment* is high because of the Chinese new year. In July 22, the number of *Technology* tweets rises abruptly because of the breakdown of WeChat

(B), which is a WhatsApp-like chat tool in China. For the left four rare categories, we find that *Sports* is fluctuated with the sports competition; for example, In May 26, sports-related tweets increase dramatically because of three sports competitions hold that day: Bayern won European Cup finals, F1 in Monaco and Beijing Guoan versus Shanghai Shenhua in Chinese Super League (C). In November 9, Hengda won the Asian Champions League matches and this made the number of sports tweets increases (D). And in June 11, *Military* has a peak as the successful launch of Shenzhou-10 (E). So the fluctuation of the seven topics is in fact a mirror of social events, and it can be used to discover important social events through probing the Weibo stream.

We also notice an interesting phenomenon that even though the fraction of *Finance* tweets is small, it demonstrates a significant cyclical characteristic. Further observations demonstrate that the drops of the curve always correspond to the weekends. And this phenomenon stimulates us to research the weekly pattern of Weibo topics. Then, we collect all tweets of a specific day of a week and calculate the seven topics' proportions at that day. The result is shown in Fig. 3 (left column). The occupations of tweets in Saturday and Sunday are apparently different from that of weekdays. At the weekend, people will pay more attention to *Entertainment*, while the ratio of *Finance* and *Technology* decline sharply. And for other topics, their weekly patterns are trivial, which are not reported here.

To explore the periodicity of Weibo topics, we also use discrete Fourier transform (DFT), a typical method in signal processing, to convert the time-domain data of Fig. 2 to frequency-domain data ( $Hz$ ). We assume that the duration from January 06, 2013 to November 30, 2013, i.e., 329 days in total, is one second. Then, the sampling rate should be 329. In other words, we sample 329 points in one second. The frequency is defined as the number of periodicity per unit time. In this condition, if the curve is strictly on seven-day cycle, the frequency would be 47. From Fig. 3b, except  $Hz = 0$ , the peaks are  $Hz = 47$  and  $Hz = 94$ . The frequency of 47 corresponds with weekly cycle, which validates our assumption, while the 94 frequency is equivalent to the 3.5-day periodicity, which is an extraordinary finding and deserves further explorations in the future. The spectrum of *Entertainment* is similar to *Technology*, and we only plot one of them in Fig. 3d. The periodicities of the two topics are similar with *Finance*, but the cyclicity is less significant since other frequencies occupy much more parts as compared to *Finance*. We also present the spectrum of *Society* to represent the four topics that possess no apparent weekly cycle in Fig. 3f. To sum up, we find that weekly cycle exists in *Finance*, *Entertainment* and *Technology*, but it vanishes in *Society*, *International*, *Military* and *Sports*, which are easily affected by the real-world events. It is worthy noting that

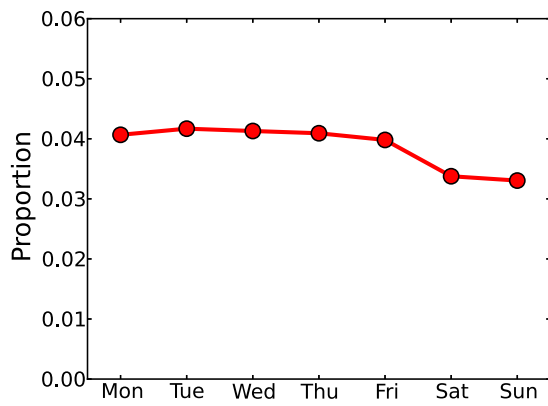
*Finance* has the most significant periodic pattern, which further justifies our previous finding.

## 4.2 Spatial pattern

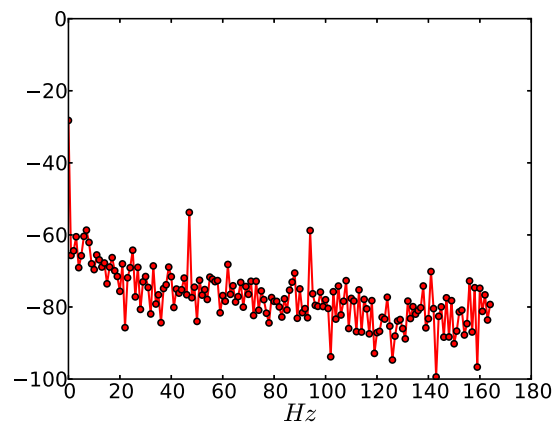
Being an essential dimension in human activity, geography might influence the topic expression in online social media. Actually, topic distribution of tweets posted at a certain place can reflect the interest of people who live there. And it might be related to the local economy or custom. It is also reasonable to assume that the registration location (at the level of province) of a user is his or her permanent place of activity, especially for those active users. Then, we can label each tweet with a geographic location based on its author's residence. To disclose the spatial pattern, we first label tweets into 34 provinces of China and then calculate the topic distribution in each of them. The results are shown in Fig. 4. From Fig. 4a, the fraction of *Entertainment* is particularly high in Hong Kong and Taiwan, which is consistent with our intuition that both places are renowned for movies stars and entertainment TV shows. With respect to the mainland, we find that the *Entertainment* proportions in Beijing, Shandong, Henan, Shanxi and Hebei are relative low, which may be accounted to the stressful life in these areas. On the contrary, these provinces possess more technology tweets than others, especially for Beijing. Intuitively, the more developed a province's economy, the more *Entertainment* tweets might be posted, but the experiment overthrows this assumption and indicates that the activity of entertainment is not strongly correlated with the level of regional income.

## 4.3 Gender pattern

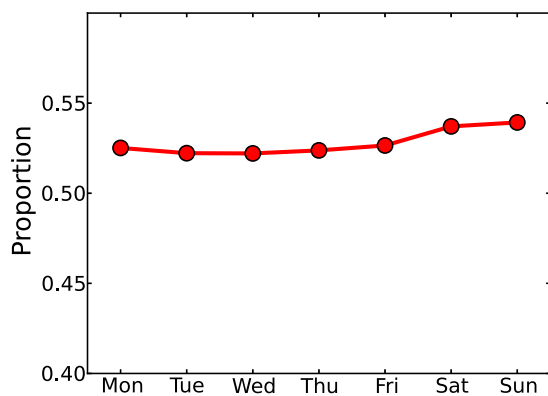
Intuitively, male and female users concentrate on different topics. Through the profile of massive users in the streaming dataset, we can validate this opinion by analyzing the enormous volume of data. The result is shown in Fig. 5. First of all, the number of tweets posted by female users is much higher than that of the male, which means the majority of active users in Weibo are female. And in different topics, the value of gender proportion is also diversified. Specifically, in *Society*, the distribution is almost the same with the entire case, while for *Entertainment*, tweets posted by female users are much more than that posted by the male and the gap is even larger than that of the entire case. However, in other topics, male users post more tweets, especially in *Military*, *Finance* and *Sports*. We then consider two important measurement named followers' count and verifying status, which represent a user's influence. Followers' count is the number of users who follow the specified user. The more followers a user owns,



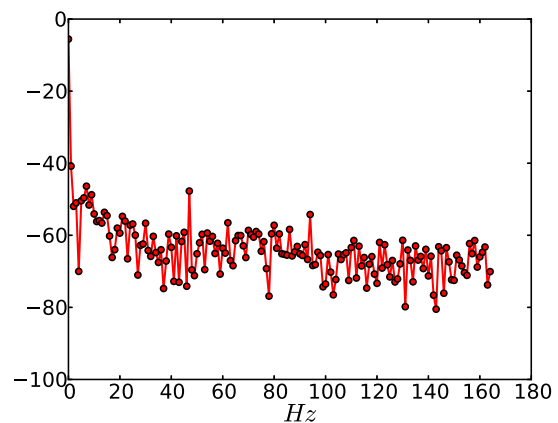
(a) Finance



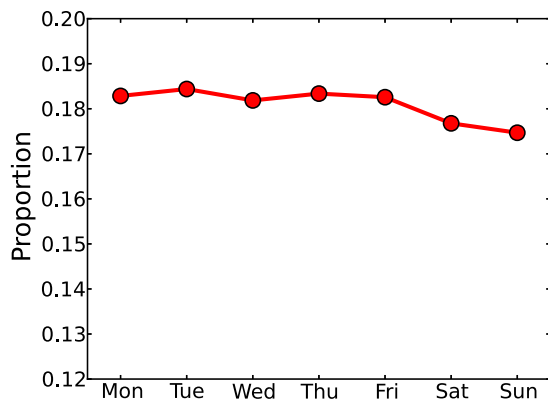
(b) Finance Spectrum



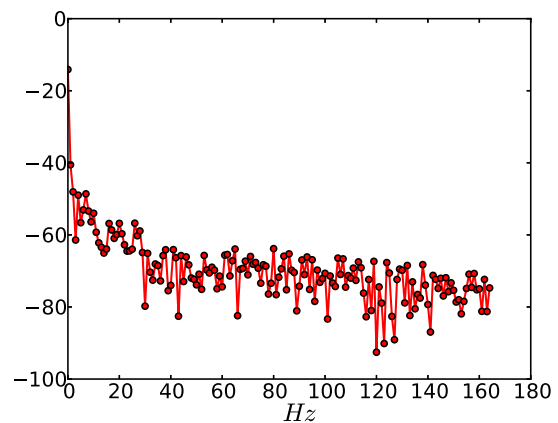
(c) Entertainment



(d) Entertainment Spectrum



(e) Technology



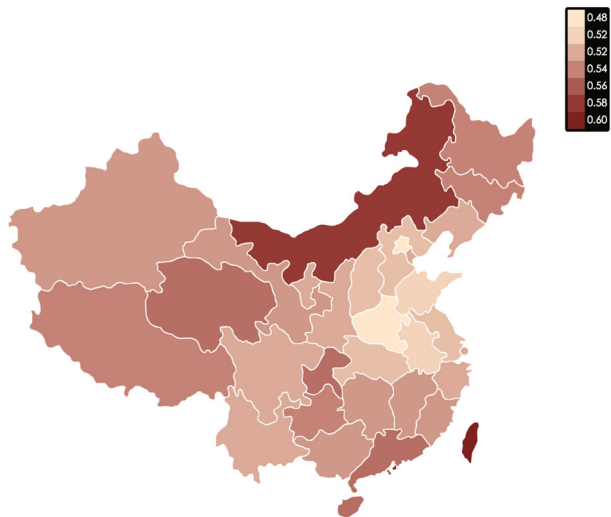
(f) Society Spectrum

**Fig. 3** The weekly pattern of Weibo topics. In the three spectrum figures (right column), the horizontal axis represents the frequency, while the vertical stands for the value in terms of decibels

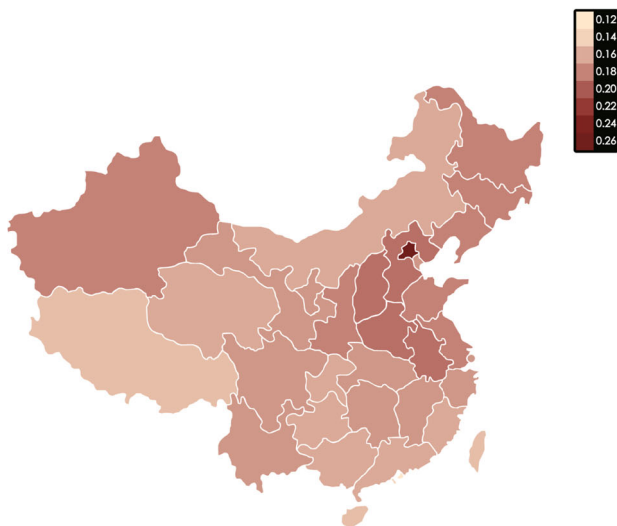
the more influence he/she possesses. The verified accounts are important users who are officially verified by Weibo. Generally speaking, verified users are well-known people who possess many followers and have relatively huge

influence. The result is depicted in Fig. 6, which shows that both of the two metrics in *Finance* are larger than that in others, while in *Entertainment* their values are particularly low. It tells us that the *Finance* tweets are always technical





(a) Entertainment



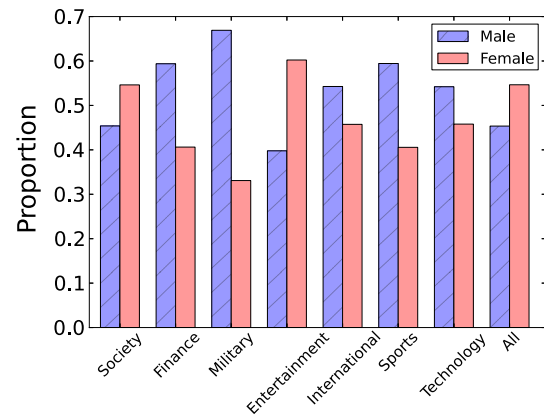
(b) Technology

**Fig. 4** The spatial pattern of Weibo topics. The *color* represents the ratio of a specified topic in this province. *Darker color* stands for more tweets of a certain topic generated by the corresponding province. The value ranges from 48 to 60 % in **a**, from 12 to 26 % in **b**

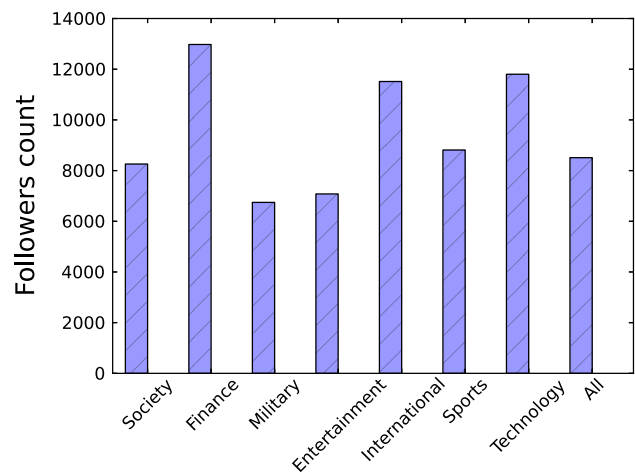
and attract more expert users. And considering the gender of topic together, we can draw a conclusion that male users tend to concentrate more on professional topics than the female in Weibo.

#### 4.4 Sentimental pattern

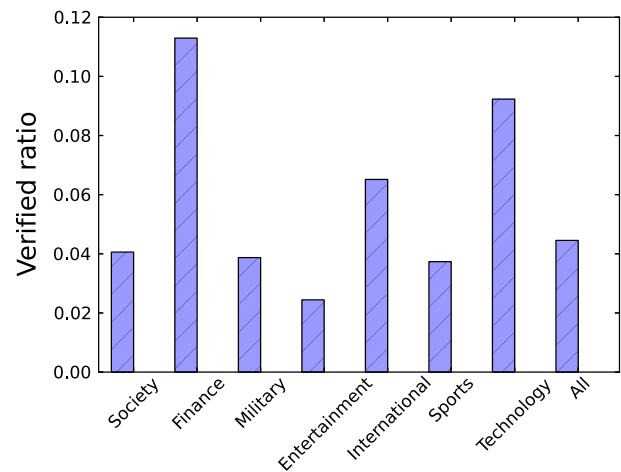
Among the sophisticated signals delivered by tweets, the sentiment is a convincing proxy of users' opinion to a



**Fig. 5** Gender distribution of each topic. Male and female users concentrate on different topics



(a) Averaged number of followers



(b) Verified ratio

**Fig. 6** **a** Comparison of averaged number of followers. **b** Comparison of ratio of verified users

certain topic. To analyze the correlation between topic and sentiment in Weibo, we employ a sentiment analysis tool named *Moodlens* (Zhao et al. 2012) to label a tweet with an emotion (anger, disgusting, joy and sadness). Combined with our topic classifier, we get the sentiment distribution of each topic, which is shown in Fig. 7. In the topic of *Society*, sad tweets is the majority. As compared to the emotion distribution of all tweets, the fraction of joyful tweets declines and both of anger and sadness raise. Moreover, angry emotion even becomes the majority sentiment in *Finance* and *Military*. In fact, it indicates that in China, two types of events easily stimulate the angry emotion in Weibo (Fan et al. 2014). The first one is the unfair social phenomenon such as the skyscraping housing price and the government corruption. These angry tweets exist in topic *Finance* and *Society*. And another one is the diplomatic conflict between China and foreign countries such as visits of the Yasukuni War Shrine and the drill of US–South Korea. And these events could be reflected in *Military* and *International*. Regarding *Entertainment* and *Technology*, joy is the dominant emotion, which is consistent with our intuition. Particularly, the fraction of joyful tweets is surprisingly higher in *Technology*, which is about 60 % in this topic. By analyzing the original data, we find tweets in this topic include many advertisements automatically posted by mobile applications, which are always joyful.

Then, we calculate the emotion entropy of each topic, from which we can learn that emotion distributions of different topics vary widely; for instance, tweets in *Finance*, *Military* and *International* have high entropy, indicating evenly distributed emotions in these topics. As for *Entertainment* and *Technology*, the emotion distributions are obviously biased, indicating the existing of dominant emotions.

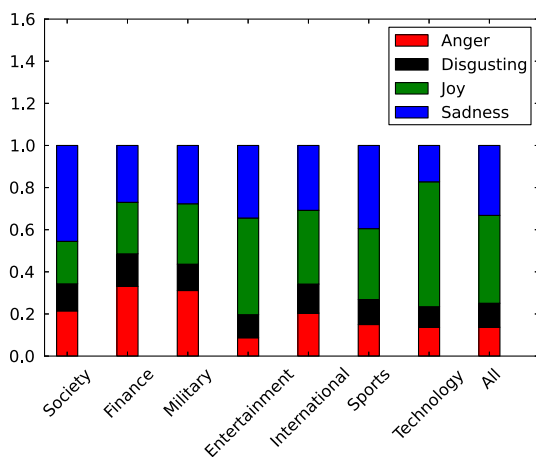


Fig. 7 Emotion distribution on each topic

#### 4.5 Subtopic pattern

The seven-class taxonomy is a comprehensive approach to study Weibo topics and for each of them, tweets can still be clustered into different semantic groups, which here we denote as subtopics; for example, *Finance* tweets may be related to the GDP, the monetary policy or the real estate market, respectively. Meanwhile, the fluctuation of the temporal curve is caused by some events such as earthquakes or the film festival, which can also be regarded as subtopics. Therefore, studying the pattern of subtopics in each topic indeed deserves further explorations.

Topic models such as PLSA (Hofmann 1999) and LDA (Blei et al. 2003) are widely used in clustering text corpus. Here we employ LDA to reveal subtopics. Firstly, we gather tweets of a specified topic and compose a corpus of tweets for each topic. Considering the shortness of the tweet and LDA's disadvantage in handling sparse texts, we slice the corpus by hour and combine tweets posted in the same hour as one document to solve the short-text problem in modeling Weibo. Based on the new corpus with merged documents, PLDA<sup>2</sup>, a parallel LDA implementation is used to mine the subtopics for each topic. Among the 100 subtopics obtained, we then manually select meaningful ones for later analysis. Surprisingly, we find that some real-world incidents appear in almost all the topics; for example, because of the huge influence of the Ya'an earthquake, the subtopic *earthquake* appear in all of the seven topics. Moreover, the subtopic *Diaoyu Island* exists in *Finance*, *Military*, *Society* and *International*. It can be interpreted that hot issues may trigger extensive discussion from different perspectives in Weibo. Taking *Diaoyu Island* incident as an example, on the one hand it is an international conflict, but on the other hand, it is also a military confrontation. Meanwhile, people can also evaluate this conflict from the view of *Finance* and *Society*. Sharing common subtopics can confuse different topics by blurring boundaries between them, so the above finding in fact empirically explains why the accuracy of topic classification can not be as high as expected. It also demonstrates the possibility that combining views of different topics helps to get a complete picture of an incident, which could be a promising direction in future works. Note that the same keyword appears in more than one subtopics might be also caused by the poor performance of the classification model, while here we perform PLDA on well-preprocessed and well-combined tweets, and therefore, its convincing performance would greatly reduce the above possibility.

<sup>2</sup> <http://code.google.com/p/plda/>.

### 4.6 Mobility pattern

The number of Internet users in China experiences a booming growth in recent decades and in which the ones with mobile devices occupy a major proportion. Until June of 2014, there are around 527 million mobile users in 632 million Chinese Internet users. It is natural to ask the question that to what extent the online social media such as Weibo is influenced by the frequent access from those mobile users; for example, the convenience of phone and tablet could definitely boost the information sharing and diffusion. Because of this, we would investigate the patterns of mobility in Weibo topics in this subsection.

Different from the former analysis, here the interaction network dataset from Weibo is used to conduct the statistics. In each tweet, we have its post source, which is a shot text such as “iPad Client” and “Samsung Galaxy Note II.” And this message can actually be used to determine from what device the tweet is posted. Firstly, we extract the number of each source text from our dataset and pick out the mostly used 30 ones. And tweets labeled in those source texts account for 90 % of all tweets. Secondly, from the 30 most popular sources, we pick out 16 ones that represent mobile devices such as “iPad,” “iPhone” and “Android.” We suppose that the 16 sources can represent most mobile-device-post tweets, and then, each tweet can be labeled into two categories: from mobile devices or not. Obviously, the proportion of tweets posted from mobile devices rises steeply these years, from 5 % in August 2009 to 50 % in January 2013. The prevalence of using phones could significantly influence the online social network, such as content and social ties. In this paper, we explore this issue from the view of Weibo topics.

First we calculate the proportion of mobile tweets in each topic and find the evident variance among different topics, which is shown in Fig. 8. Note that the proportion of *Society* tweets from mobile devices is particularly high, indicating that mobile users concentrate more on social events. Besides, the fractions of *Finance* and *Technology* are surprisingly lower than the average, suggesting the declining interest on both topics for mobile users. The reason of these phenomena is that the entertainment-oriented function of the usage of mobile devices. To be specific, when people are attracted by the wonderful content on their mobile phones, they always tend to browse information that are relative non-professional, such as social news and refuse technical materials such as finance or technology affairs.

### 4.7 Retweeting pattern

Being a typical feature in Twitter and its variants, retweeting is the most popularly way to diffuse tweets and

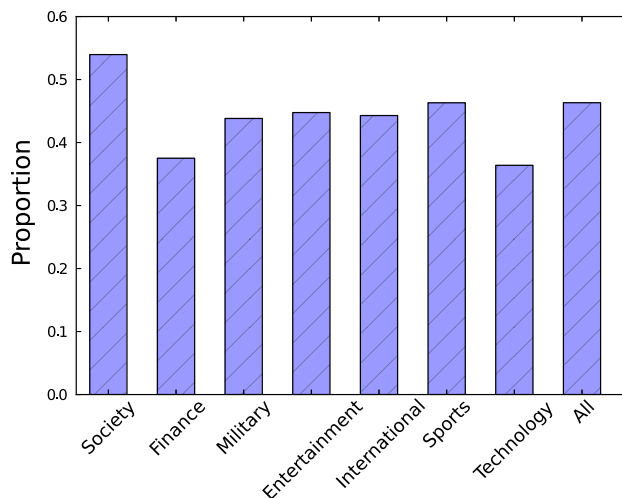


Fig. 8 The proportion of tweets posted from mobile devices in each topic

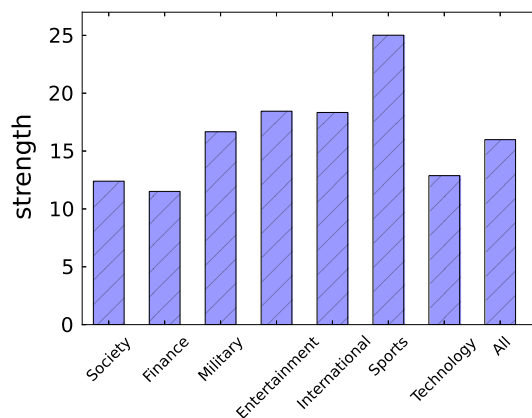


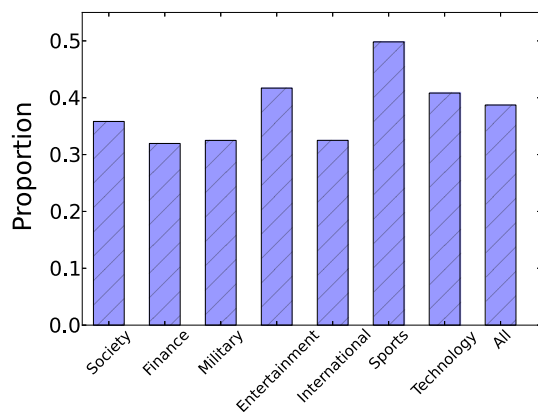
Fig. 9 RPS of the seven topics. *Sports* tweets always propagate through close friends

spread opinions. By constructing an interaction network, in which each node stands for a user, the link stands for the retweeting action between them and the number of retweeted tweets stands for the link weight, we can investigate the retweeting patterns in different topics of Weibo in this subsection. Note that here only the tweets of 2013 in the second dataset are considered, and meanwhile, the network itself can also be redefined by the link threshold  $T$ , which means the retweeting number on each link should be equal or greater than  $T$ .

For each topic, we first compare their retweeting path strength (RPS), which is defined as the average link weight between pairs of users retweeting the correspond topic. In fact, RPS can reflect the strength of social ties on which the retweeting occurs. In order to keep the weak ties in network, we let  $T = 1$  and the results are plotted in Fig. 9. As can be seen, we find that RPS of *Sports* is particularly high,

while RPS in *Society*, *Finance* and *Technology* is relatively low. It indicates that Weibo users tend to share *Sports* messages with their close friends, while the latter three topics are comparatively technical and inclines to diffuse through weak ties.

From the perspective of social network, strong ties mainly distribute inner-community and weak ties distribute inter-communities. The difference in RPS preference makes the discussion of retweeting topics on the community-level worthwhile. Through the community detection approach Infomap (Rosvall and Bergstrom 2008), we can calculate the ratio of inner-community retweetings on each topic in the interaction network. To improve the quality of community detection algorithm, we set  $T$  as 10 to delete occasionally retweetings and an interaction network with 20,576 nodes, 45,643 links and 27,960,329 tweets posted in



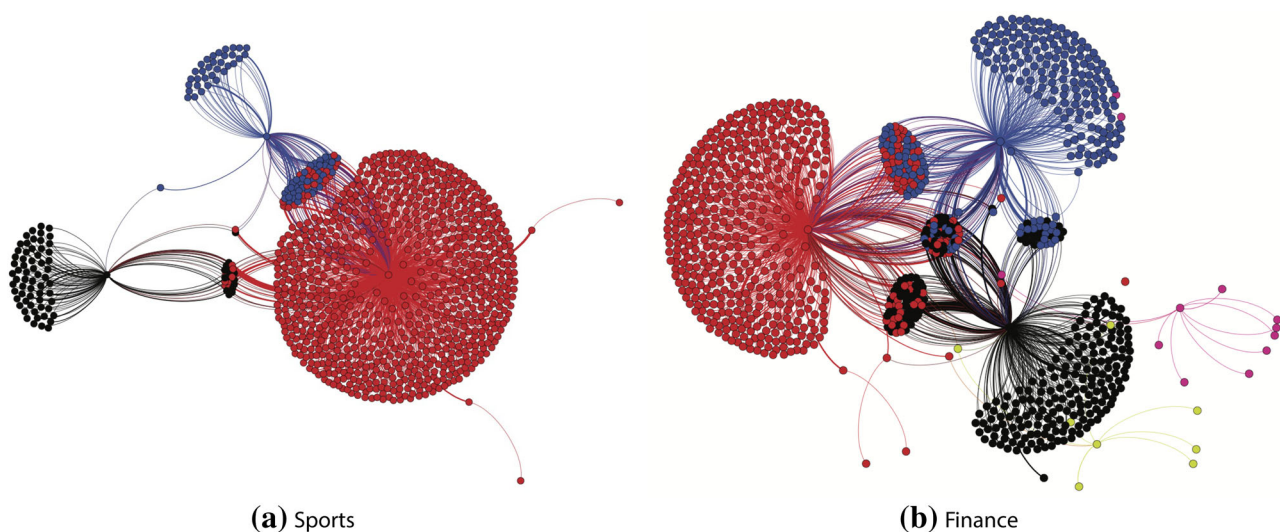
**Fig. 10** The ratio of inner-community retweeting behavior in seven topics

2013 is constructed. As shown in Fig. 10, *Sports*, *Technology* and *Entertainment* always propagate in a closeness community, while *Society*, *Finance*, *Military* and *International* tend to diffuse across different communities. Consistent results are also demonstrated in Fig. 11, in which the topology of interaction networks for *Sports* and *Finance* is compared. It can be seen that *Sports* possesses much more inner-community retweetings, while *Finance* shows more inter-communities exchanges.

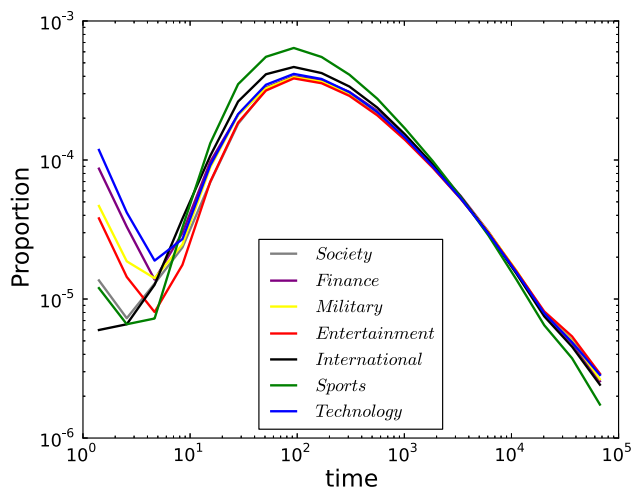
Next, we discuss the retweeting interval of each topic, which reflects the velocity of the information diffusion. The interval of a tweet is defined as the length of time after which it is retweeted, and here we only investigate the intervals less than 24 h. The distribution of seven topics' retweeting intervals is depicted in Fig. 12. We find that all the curves show power-law-like declines as the interval is higher than 100 s. It is consistent with previous studies on human dynamics (Barabasi 2005; Zhou et al. 2008). We also notice that the retweeting interval of the seven topics is different, as listed in Table 3, and the *Sports* messages is always reposted quickly. It is consistent with our former observation that *Sports* messages are preferably delivered among close friends.

#### 4.8 Correlation pattern

Tweets of different topics might attract different users, and the same user can focus on more than one topic simultaneously. Hence, through mining the relevance of different topics, we can learn the spectrum of users' interests. Similarly, for each topic  $i$  an interaction network can be established, which is denoted as  $G_i(V_i, E_i, T_i)$ , where  $V_i$  is the node set,  $E_i$  is the link set, and the weight threshold



**Fig. 11** Comparison of interaction network topologies. Several typical communities in each network are demonstrated, and *different colors* denote different communities. For each community, we only keep the ties with retweeting times larger than 10 for better visualization



**Fig. 12** The distribution of retweeting interval for seven topics. The intervals is limited in 24 h

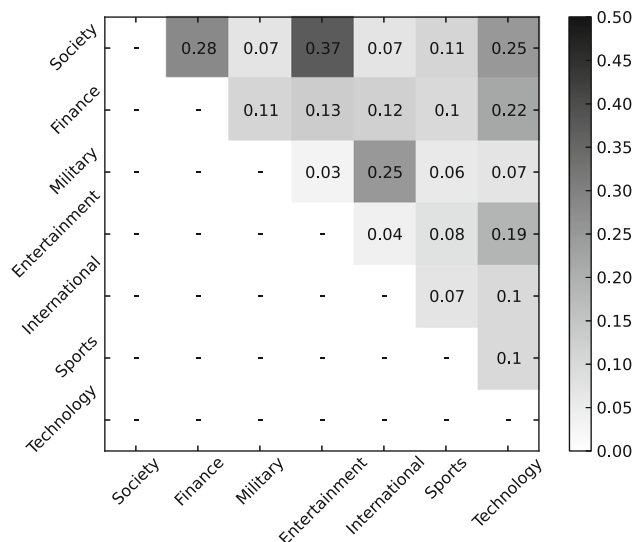
**Table 3** The average retweeting interval of the seven topics

Topic	Interval
Society	15,536.86
Finance	14,457.40
Military	14,906.77
Entertainment	15,858.98
International	13,771.21
Sports	<b>11,063.17</b>
Technology	15,286.60

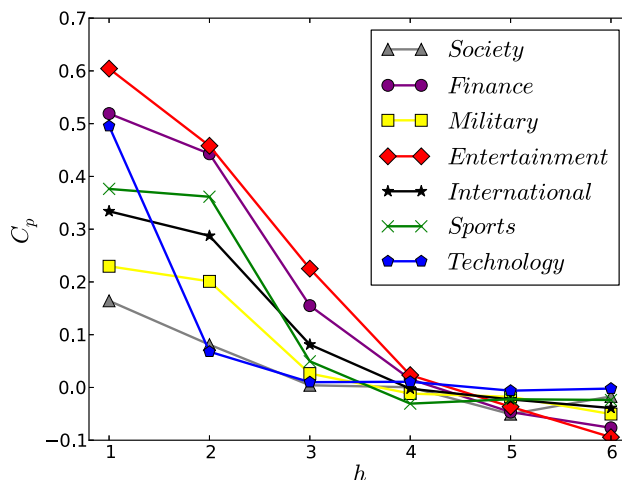
Bold indicates the minimum value

$T_i = 5$ . We think that users in the graph  $G_i$  are ones that focus on topic  $i$  since they participate in this topic for more than  $T$  times. Then, we calculate the similarity  $S_{ij}$  of a pair of topics  $i$  and  $j$  with their corresponding user set  $V_i$  and  $V_j$  through  $\frac{V_i \cap V_j}{V_i \cup V_j}$ . As shown in Fig. 13, the similarity of *Entertainment* and *Society* is the highest, which reaches 0.37, indicating a strong relevance. Besides, similarity of *International* and *Military* is 0.25, suggesting a correlation that is consistent with our expectation, while for *Sports*, it rarely shows relevance with other topics.

The correlation of Weibo topics can also be probed by directly measuring the topic similarity of different users in the interactive network. In this experiment, we set  $T = 10$  to guarantee the high link strength. Firstly, a vector that contains the proportions of topics is generated for each user, which could be denoted as  $e_i(f_1^i, f_2^i, \dots, f_7^i)$ . Then, for a specific hop  $h$ , we pick out all user pairs with distance  $h$  from the network. Secondly, for a given user pair  $(u_i, u_j)$  and a fixed  $h$ , we put  $u_i$ 's  $f_m^i$  into a sequence  $S_m$ , and  $u_j$ 's  $f_m^j$  to another sequence  $T_m$ , where  $m$  is an arbitrary topic. Then, Pearson correlation coefficient of



**Fig. 13** The similarity of all topic pairs



**Fig. 14** Correlations of the seven topics. The horizontal axis represents the social distance of users, while the vertical axis represents the Pearson correlation of each topic

$S_m$  and  $T_m$  will be calculated to reflect the strength of correlation between different users for topic  $m$  as the social distance is  $h$ . From Fig. 14, we find that the correlation indeed exists for all the topics and almost all of their strengths are greater than 0.2 when  $h = 1$ , which indicates that connected users in Weibo tend to share similar topics through frequent interactions. Meanwhile, the correlations drop dramatically as  $h$  grows, which is consistent with the finding in Fan et al. (2014). In fact, since  $h > 3$ , the correlation of all topics is almost zero. Although occupying a small fraction of tweets, *Finance* and *Technology* still demonstrates high correlations as  $h = 1$ , which further indicates the great influence of these two topics.



## 4.9 Patterns summary

To sum up, significant patterns disclosed above are summarized as follows.

*Entertainment* is the absolutely majority in Weibo, which is significantly different with that in Twitter. Specifically, Twitter users pay more attention on *Political* affairs, while Chinese users are always attracted by *Entertainment* messages. Topics that Weibo users concern are related to present hot events, which can be used to monitor real-world events. Besides, topics' periodic pattern is also revealed and *Finance* shows the most significant cyclicity.

With the help of *Moodlens*, a sentiment analysis tool, we unravel the correlation between topics and sentiments. The result shows that anger is the general response for *Finance* and *Military* messages, while joy sentiment dominates *Entertainment* tweets.

By analyzing the repost paths, we reveal retweeting patterns in three angles: the RPS, the community and the retweeting interval. Weibo users tends to share *Sports* messages with their close friends, while *Finance* information always diffuses through weak ties and shows more inter-communities exchanges.

Topics might be concentrated by the same users, such as *Entertainment* and *Society*, *Military* and *International* and then demonstrate correlations between them. Homophily is also revealed in Weibo topics which means connected users tend to share similar topics through frequent interactions; for example, *Entertainment* and *Finance* possess highest correlations when  $h = 1$ .

## 5 Conclusion and future work

Recent years have witnessed the rapid growth of the online social media in China, and it has become the largest platform for users to post their statuses and opinions everyday. Considering the massive users and large amount of content, online social media indeed provide a unprecedented big data window to probe human behaviors, in which the topic expression and preference is an essential one. In this paper, by building a fast and self-updating topic classifier, we comprehensively investigate topic dynamics of Weibo through two datasets. Plenty of interesting patterns are then revealed from different views, including time, geography, demographics, sentiment, subtopics, social mobility, retweeting and correlation.

Our findings could shed light on many applications about topics in online social media; for example, temporal patterns can help the development of event detection and tracking solution, while the difference existing in provinces and gender might help design location-flexible and gender-aware strategies for content recommendation methods. The

fact of different topics sharing subtopics suggests that a multi-topics view could provide a new and complete picture in understanding real-world events. And different preferences in selecting tie strengths inspire us to explore the diffusion discrepancy for different topics in future work.

**Acknowledgments** This work was supported by NSFC (Grant No. 61421003) and the fund of the State Key Lab of Software Development Environment (Grant No. SKLSDE-2015ZX-05). Jichang Zhao was partially supported by the fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2015ZX-28) and the Fundamental Research Funds for the Central Universities (Grant No. YWF-15-JGXY-011).

## References

- Ardon S, Bagchi A, Mahanti A, Ruhela A, Seth A, Tripathy RM, Triukose S (2013) Spatio-temporal and events based analysis of topic popularity in Twitter. In: Proceedings of the 22nd ACM international conference on conference on information & knowledge management (CIKM), San Francisco, CA. ACM, pp 219–228
- Banerjee S, Ramanathan K, Gupta A (2007) Clustering short texts using Wikipedia. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp 787–788
- Barabasi AL (2005) The origin of bursts and heavy tails in human dynamics. *Nature* 435(7039):207–211
- Becker H, Naaman M, Gravano L (2011) Beyond trending topics: real-world event identification on Twitter. In: ICWSM
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Bogdanov P, Busch M, Moehlis J, Singh AK, Szymanski BK (2013) The social media genome: modeling individual topic-specific behavior in social media. In: Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining. ACM, pp 236–242
- Boyd D, Golder S, Lotan G (2010) Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. In: System sciences (HICSS), 2010 43rd Hawaii international conference. IEEE, pp 1–10
- Cataldi M, Di Caro L, Schifanella C (2010) Emerging topic detection on Twitter based on temporal and social terms evaluation. In: Proceedings of the tenth international workshop on multimedia data mining. ACM, p 4
- Dumais S, Platt J, Heckerman D, Sahami M (1998) Inductive learning algorithms and representations for text categorization. In: Proceedings of the seventh international conference on Information and knowledge management. ACM, pp 148–155
- Fan R, Zhao J, Chen Y, Xu K (2014) Anger is more influential than joy: sentiment correlation in Weibo. *PLoS One* 9:e110, 184
- Genc Y, Sakamoto Y, Nickerson JV (2011) Discovering context: classifying tweets through a semantic transform based on Wikipedia. In: Foundations of augmented cognition. Directing the future of adaptive systems. Springer, pp 484–492
- Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp 50–57
- Joachims T (1999) Transductive inference for text classification using support vector machines. In: ICML, vol 99, pp 200–209

- Kinsella S, Passant A, Breslin JG (2011) Topic classification in social media using metadata from hyperlinked objects. In: *Advances in information retrieval*. Springer, pp 201–206
- Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: *Proceedings of the 19th international conference on World Wide Web, WWW '10*. ACM, pp 591–600
- Michelson M, Macskassy SA (2010) Discovering users' topics of interest on Twitter: a first look. In: *Proceedings of the fourth workshop on analytics for noisy unstructured text data*. ACM, pp 73–80
- Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, pp 29–42
- Novakovic J (2010) The impact of feature selection on the accuracy of naïve Bayes classifier. In: *18th telecommunications forum TELFOR*, pp 1113–1116
- Quercia D, Capra L, Crowcroft J (2012) The social world of Twitter: topics, geography, and emotions. In: *ICWSM*
- Ritter A, Etzioni O, Clark S et al (2012) Open domain event extraction from Twitter. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp 1104–1112
- Romero DM, Meeder B, Kleinberg J (2011) Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In: *Proceedings of the 20th international conference on World Wide Web*. ACM, pp 695–704
- Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci* 105(4):1118–1123
- Sankaranarayanan J, Samet H, Teitler BE, Lieberman MD, Sperling J (2009) Twitterstand: news in tweets. In: *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM, pp 42–51
- Schönhofen P (2009) Identifying document topics using the Wikipedia category network. *Web Intell Agent Syst* 7(2):195–207
- Song S, Li Q, Bao H (2012) Detecting dynamic association among Twitter topics. In: *Proceedings of the 21st international conference companion on World Wide Web*. ACM, pp 605–606
- Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M (2010) Short text classification in Twitter to improve information filtering. In: *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval*. ACM, pp 841–842
- Suh B, Hong L, Pirolli P, Chi EH (2010) Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network. In: *Social computing (socialcom), 2010 IEEE second international conference*. IEEE, pp. 177–184
- Yamaguchi Y, Amagasa T, Kitagawa H (2011) Tag-based user topic discovery using Twitter lists. In: *Advances in social networks analysis and mining (ASONAM), 2011 international conference*. IEEE, pp 13–20
- Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: *Proceedings of the fourth ACM international conference on web search and data mining*. ACM, pp 177–186
- Yang T, Lee D, Yan S (2013) Steeler nation, 12th man, and boo birds: classifying Twitter user interests using time series. In: *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*. ACM, pp 684–691
- Yang Y, Liu X (1999) A re-examination of text categorization methods. In: *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*. ACM, pp 42–49
- Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: *ICML*, vol 97, pp 412–420
- Yang Z, Guo J, Cai K, Tang J, Li J, Zhang L, Su Z (2010) Understanding retweeting behaviors in social networks. In: *Proceedings of the 19th ACM international conference on information and knowledge management*. ACM, pp 1633–1636
- Yu L, Asur S, Huberman BA (2011) What trends in Chinese social media. In: *The 5th SNA-KDD workshop'11 (SNA-KDD'11)*, 21 August 2011, San Diego, CA
- Yu L, Asur S, Huberman BA (2015) Trend dynamics and attention in Chinese social media. *Am Behav Sci*. doi:[10.1177/0002764215580619](https://doi.org/10.1177/0002764215580619)
- Zhang T, Oles FJ (2001) Text categorization based on regularized linear classification methods. *Inf Retr* 4(1):5–31
- Zhao J, Dong L, Wu J, Xu K (2012) Moodlens: an emoticon-based sentiment analysis system for Chinese tweets. In: *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp 1528–1531
- Zhou T, Han XP, Wang BH (2008) Towards the understanding of human dynamics. In: *Science matters: humanities as complex systems*, pp 207–233