J Syst Sci Syst Eng (Jun 2014) 23(2): 212-230 ISSN: 1004-3756 (Paper) 1861-9576 (Online)
DOI: 10.1007/s11518-014-5243-z CN11-2983/N

TOPICS AND TRENDS OF THE ON-LINE PUBLIC CONCERNS BASED ON TIANYA FORUM*

Lina Cao Xijin Tang

Academy of Mathematics and Systems Science

Chinese Academy of Sciences, Beijing, 100190, China

{caolina, xjtang}@amss.ac.cn (✉)

Abstract

Many social events spread fast through the Internet and arouse wide community discussions. Those on-line public opinions emerge into diverse topics along the time. Moreover, the strength of the topics is fluctuating. How to catch both primary topics and trend of topics over the shifting on-line discussions are not only of theoretical importance for scientific research, but also of practical importance for societal management especially in current China. To try the cutting-edge text analytic technologies to deal with unstructured on-line public opinions and provide support for social problem-solving in the big data era is worth an endeavour. This paper applies dynamic topic model (DTM) to explore the changing topics of new posts collected from Tianya Zatan Board of Tianya Club, the most influential Chinese BBS in mainland China. By analysis of the hot and cold terms trends, we catch the topics shift of main on-line concerns with illustrations of topics of *school bus* and *environment* in December of 2011. An algorithm is proposed to compute the strength fluctuation of each topic. With visualized analysis of the respective main topics in several months of 2012, some patterns of the topics fluctuation on the board are summarized.

Keywords: Topic models, dynamic topic model, on-line topics evolution, Tianya Club, societal management

1. Introduction

Social media, such as blog, microblog, review sites, BBS, etc., are fundamentally changing the way people communicate (Wu, Sun & Tan, 2013). In China, more and more people treat social media as one way to express their opinions toward the daily phenomena and social events openly and freely. The on-line discussions which show fresh, diverse and evolving opinions bring

* This research is supported by National Basic Research Program of China under Grant No. 2010CB731405 and National Natural Science Foundation of China under Grant No.71171187 & 71371107. This paper is an extended version of the paper presented at the 14th International Symposium on Knowledge and Systems Sciences (KSS2013), held in Ningbo during October 25-27, 2013 .

great influence toward societal, political and personal life. For example, in June of 2011 the name of "Guo Meimei" was very famous in Sina Weibo. The show-off of luxurious lifestyle by the 20-year-old girl who claimed close relations with Chinese Red Cross incurred the trust crisis of Chinese Red Cross. This event exerted negative effects toward the donation after Lushan earthquake happened in April of 2013 compared with those after Wenchuan earthquake happened in May of 2008. There are many similar cases, such as "My dad is Li Gang" event (occurred on October 16, 2010) and "Xiao Yueyue" event (occurred on October 5, 2010), where the foci of public opinions shifted from the cases to the institution, governance or the morality of whole society (Cui, He & Liu, 2013). Public attentions to those events are also changing. Many things suddenly happened incur the hot discussions while cool down slowly as time goes on.

Released by social media and often accumulate lots of on-line concerns, many on-line highlights refer to those serious social problems such as corruption, education inequity, environment pollution, land planning and development, national security, etc. which are actually wicked problems (Rittel & Webber, 1973), and affect social stability in current China. For societal management, how to acquire a structured vision from those emerging diverse and unstructured on-line opinions is required to deal with those wicked problems especially in Web 2.0 and big data era. While understanding the on-line public opinions and detecting the hot topics from the massive textual data are novel challenges for governance. Some computing methods, including text mining, behavior analysis, graph-based modeling, etc., provide potential solutions toward on-line opinion mining. Even it is a matter of creativity to devise practical solutions, and a matter of judgment to determine which are effective, validated and worth being implemented. Text analytics algorithms and techniques are increasingly being developed. Typical examples include document classification, document clustering, topic detection and modelling, and opinion mining/sentiment analysis. At the same time, successful applications of those innovative technologies raise a number of pertinent research questions, with both theoretical and practical ramifications.

This paper explores the unstructured on-line opinions via text analysis and modeling of posts from the biggest Chinese BBS, Tianya forum (Zhang & Tang, 2011). In this forum, new posts published in the Tianya Zatan board (TYZT board) are crawled and analyzed to detect the dynamic topics and the evolution of words of the topics. Although the microblog, such as Sina Weibo in China, is very popular in recent 3 years, Tianya forum where happened wide-scope discussions of social events and contributed a lot of insights has a longer history and enables people to undertake more in-depth discussions. Compared to blog, the on-line forum provides more interactive conversations on a particular topic.

In this paper, we study the fluctuation of topics and the evolution of words by a statistical modeling perspective. The paper is organized as follows. Firstly, relevant work is reviewed. Two models, dynamic topic model (DTM) and continuous dynamic topic model (cDTM), are briefly introduced. Secondly, the data set is described, and the fitter model is selected according to the sparsity of data. Thirdly, the analyses of modeling results are addressed. Finally, conclusions are given.

2. Relevant Methods to Analyze the Topic Evolution

People have been convinced of the importance of identifying and tracking topics since the earlier Topic Detection and Tracking (TDT) study by Allan et al. (1998). In TDT, a topic is defined to be a set of news stories that are strongly related by some seminal real-world events. The goal of TDT is to monitor the stories of those events that have not been seen before, and to gather the stories into groups. The traditional technologies in information retrieval, information management and data mining are considered to resolve the task (Wayne, 2000). But the time information of the text is not used efficiently in TDT.

Considering time information for the task of topics identification and tracking with time-stamped text data is the focus of recent studies in machine learning field (Cao, et al., 2007; Guha, et al., 2005; Kleinberg, 2002). Topic models, as statistical models, have been developed for this task. As Blei & Lafferty (2007) have been pointed out, topic modeling has become a powerful tool for "extracting surprisingly interpretable and useful structure without any explicit 'understanding' of the language by computer".

2.1 Topic Models Review

According to post-discrete, pre-discrete and continuous time, topic models can be generally divided into three categories. The post-discrete time model is the basic topic model, i.e. latent Dirichlet allocation (LDA) model, which treats documents as *bags of words* generated by one or more topics (Blei, Ng & Jordan, 2003). This model is applied to the whole documents to induce topics sets and then classify the subsets according to the time of documents (Hall, Jurafsky & Manning, 2008). A linear trend analysis on document-topic level is conducted to find topics which show statistically significant increasing or decreasing linear trend.

The pre-discrete time model, such as dynamic topic models (DTM) (Blei & Lafferty, 2006) and online LDA (OLDA) model (Alsumait, Barbara & Domeniconi, 2008), marks documents according to the discrete time before the generative process.

The continuous time model includes Topics over Time (TOT) model which captures both word co-occurrence and localization in continuous time (Wang & McCallum, 2006) and the continuous dynamic topic model (cDTM) which replaces the discrete state space model of the DTM with continuous generalization, i.e. Brownian motion (Wang, Blei &

Heckerman, 2008). Another model proposed by Nallapati, et al. (2007) called multiscale topic tomography model (MTTM) analyzes the evolution of topics at various time-scales of resolution, allowing the user to zoom in and out of the time-scale.

Many researches have flexibly applied above mentioned basic methods and models and got the desirable results. He, et al. (2009) proposed an iterative topic evolution learning framework by adopting LDA model to the citation network and developed a novel inheritance topic model. Zhang and Li (2012) compared two methods of topic evolution based on global and local documents by applying LDA model with recent 5-year NPC&CPPCC[1] news reports. Song, Lin and Tseng (2005) predicted human behaviors of receiving and disseminating emails by combining the contacts analysis and content analysis based on LDA model.

2.2 BBS Topic Evolution Analysis Review

Above-mentioned methods are being successfully applied to explore and predict the underlying structure of a variety of textual data, such as research papers (Alsumait, Barbara & Domeniconi, 2008), newswire articles (Wang, Blei & Heckerman, 2008), personal emails (Wang & McCallum, 2006) and movie synopsis (Meng, Zhang & Guo, 2012).

When facing massive on-line data which are quite different from scientific papers or news, how to choose fit methods to draw the topics with their evolutions is a primary issue. There are many relevant studies. Some focus on the Chinese BBS hot topic mining using clustering approach such as k-means clustering (You, et al, 2005), fuzzy clustering (Lu, Yao & Wei, 2008), or combined multiple clustering methods (Tang & Chen, 2010).

TDT model is also quite often applied to detecting and tracking the special events which draw intensive attention from netizens and play an important role in capturing public opinions through BBS data (Hao & Hu, 2010; Yang, Pierce & Carbonell, 1998).

In this paper, we explore to use topic models to discover the hot topics and compute the strength of the topics in Chinese BBS.

2.3 Dynamic Topic Model

Dynamic topic models (DTM) suppose the data are divided by time slice. We use following terminologies and notations to describe the data, latent variables and parameters in the DTM.

- Per-document topics. Each document is a mixture of topics and the different structures produce heterogeneous documents. Let α_t denote the per-document topic distribution at time t.
- Topics. A topic β is a distribution over the vocabulary. Let $\beta_{k,t}$ denote the word distribution of topic k in slice t. The time-series topics are modeled by a logistic normal distribution of $\beta_{k,1} \rightarrow \beta_{k,2} \rightarrow \cdots \rightarrow \beta_{k,T}$.
- Topic proportions. Let $\theta_{d,t}$ denote the topic distribution for document d in time t and η is the log proportions of $\theta_{d,t}$.

[1] NPC is the abbreviation of National People's Congress. CPPCC is the abbr. of Chinese People's Political Consultative Conference.

• Topic assignments. Each word is assumed drawn from one of the K topics. Let $z_{t,d,n}$ denote the topic assignment for the nth word in document d at time t.

• Words and documents. The only observable random variables are *words* which are organized into documents. Let $w_{t,d,n}$ denote the nth word in the dth document at time t.

In DTM, the multinomial distributions α_t and $\beta_{k,t}$ are generated from $\alpha_{t\text{-}1}$ and $\beta_{k,t\text{-}1}$, respectively. The generative process for slice t of a sequential corpus is as follows (Blei & Lafferty, 2006):

1. draw topics $\beta_t|\beta_{t\text{-}1} \sim N(\beta_{t\text{-}1}, \sigma^2 I)$.
2. draw $\alpha_t|\alpha_{t\text{-}1} \sim N(\alpha_{t\text{-}1}, \delta^2 I)$.
3. for each document:

(a) draw $\eta \sim \mathrm{N}(\alpha_t, a^2 I)$.

(b) for each word:

i. draw $Z \sim Mult(\pi(\eta))$.

ii. draw $\mathrm{W}_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$.

Note that π maps the multinomial natural parameters to the mean parameters,

$$\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})} \quad (1)$$

$$\theta = \pi(\eta) = \frac{\exp(\eta)}{\sum_i \exp(\eta_i)} \quad . \quad (2)$$

In DTM, how to learn the parameters according to the only observable $W_{t,d,n}$ constitutes an inference problem. Blei and Lafferty (2006) considered Gibbs sampling for inference was more difficult than that in static models, due to the nonconjugacy of the Gaussian and multinomial distributions. They then applied variational methods, in particular, the variational Kalman filtering and the variational wavelet regression for inference.

2.4 Continuous Dynamic Topic Model

In the DTM, one parameter controls the variance at each time slice. As the resolution gets finer, parameter multiplies at the cost of time and memory to support the posterior inference. In the cDTM, the variance is "a function of the lag between observations, and the probabilities at discrete steps between those observations need not be considered" (Wang, Blei & Heckerman, 2008).

A sparse variational inference is proposed to handle this model (Wang, Blei & Heckerman, 2008). Compared to DTM, cDTM introduces more latent variables without sacrificing memory or speed. Actually, this seemingly more complicated model is simpler and more efficient to fit different granularities. In this paper, the sparsity analysis is employed for fast model comparison.

3. BBS Data Processing and Analysis

Tianya Club has been once the globally biggest Chinese Internet forum whose approximately 91% visitors come from China mainland and has become a comprehensive virtual community including online forums, social media and blogging. Among a variety of BBS boards, TYZT board, the 2nd largest board within the Club, is a specific board including posts covering a wide scope of topics on daily lives, social unfair, corruption, phenomena of society, etc. Posts on major social events are always

published there and then change into hot posts with large number of clicks and replies. For example, *Sun Zhigang case*[2] empowered the public to challenge the authorities of local government and facilitated revisions of legislation (Jin, 2008).

In order to study on-line discussions in current China society, we start to collect data from TYZT board and several other relevant boards since October of 2010 (Zhang & Tang, 2011). Now there are around 2,000 new posts published and 4,000 plus posts updated every day. In this paper, we test topic modeling to the posts of this board.

3.1 Data Processing

The basic data are original posts (the first post of each new thread) from December of 2011 to December of 2012. The 13-month data are cleaned respectively as the data set. Firstly, the posts with urls but no contents are removed. The amounts of posts in each month are listed in Table 1. Secondly, the articles are segmented to words using ICTCLAS[3]. Here only nouns and gerunds are selected. They constitute the corpora and the total frequencies are listed in 3rd column of Table 1. Next, the selected terms which occur fewer than 50 times and in fewer than 10 posts are removed. The pruned words make up the dictionaries and the amount of those unique words are listed in 4th column of Table 1.

Table 1 The statistics of post data sets

Time span	Original posts #	Corpus # (thousand)	Diction -ary #
Dec. 2011	12,155	1700	4,541
Jan. 2012	12,032	1471	3,973
Feb. 2012	20,124	2700	6,091
Mar. 2012	37,549	5025	9,516
Apr. 2012	32,939	3976	8,089
May. 2012	33,471	4074	8,105
Jun. 2012	24,371	2844	6,097
Jul. 2012	30,657	3438	7,175
Aug. 2012	40,231	4276	8,299
Sep. 2012	37,418	3916	7,623
Oct. 2012	39,158	4330	8,579
Nov. 2012	42,100	4425	8,459
Dec. 2012	40,527	4635	8,603

3.2 Sparsity of Data

The evolution of topics in the forum infers that the terms at each time stamp are changing along the time. It means that not all vocabulary words are used at each measured time stamp. How much the terms change at each stamp requires a preliminary measure to help us choose fit and efficient model. The natural sparsity of text is used to measure the data set.

The sparsity of the data set is defined by Wang, Blei & Heckerman (2008):

$$\text{Sparsity} = 1 - \frac{\sum_t \sum_w \delta_{t,w}}{VT} \quad . \tag{3}$$

In Equation (3), $\delta_{t,w}$ is a 0-1 variable, and

[2] Sun Zhigang was a young man came from Hubei and worked in Guangzhou in March of 2003. Due to forgetting the ID card and temporary living permit, he was detained by police on March 20. Three days later, he was beat to death in the detention center. This case facilitates to the end of the custody and repatriation regulation in 2003.

[3] ICTCLAS is a widely used Chinese segmentation program. The website is http://www.ictclas.org/ .

$\sum_t \sum_w \delta_{t,w}$ is the sum of the number of unique terms at each time point. V is the size of the vocabulary. T is the number of the time stamps. Higher sparsity value indicates a sparser data set, or a term appears at fewer total time points.

We consider the sparsity of those 1-month data sets as listed in Table 1, together with two longer lengths, the Jan-March of 2012 and Jan-June of 2012 data sets. They are separated respectively by 1-day and 3-day. Table 2 lists the sparsity of each data set with different time stamps.

Table 2 Sparsity of monthly post data

Data set	Sparsity (1-day)	Sparsity (3-day)
Dec. 2011	0.17	0.04
Jan. 2012	0.17	0.04
Feb. 2012	0.16	0.03
Mar. 2012	0.16	0.04
Apr. 2012	0.18	0.12
May. 2012	0.18	0.05
Jun. 2012	0.23	0.13
Jul. 2012	0.20	0.05
Aug. 2012	0.18	0.05
Sep. 2012	0.17	0.12
Oct. 2012	0.18	0.05
Nov. 2012	0.16	0.12
Dec. 2012	0.17	0.05
Jan.-Mar.	0.42	0.33
Jan.-Jun.	0.71	0.29

In Table 2, the sparsity of data in each month of 2012 is about 0.2 as the time stamp is 1-day. It means that a term appears around 80% of the total time points, indicating the terms on each day are quite similar. When the time span gets wider, such as 3-month, from January to March, the sparsity value increases to 0.42. It means that the differences of terms become larger. High value indicates the cDTM is more efficiency over the DTM. Here, a rather low value of sparsity indicates DTM is more fit. Thus DTM is selected to analyze the topics evolution over posts of TYZT board.

4. DTM Applications

In DTM, discrete time slice is used to divide the data. We model the forum posts of each slice with a K-component topic model, where the topics associated with slice t evolve from the topics associated with slice t-1. How to decide the granularity of time is a quite important but confused task. A series of experiments with different intervals, 1-day, 3-day and 7-day, are attempted by running through DTM package[4] for comparisons; the results are analyzed using R. It seems that it is not suitable to train our data with too long intervals. Unlike the topics toward research articles where the new papers are based on the existing research with good consistency, the BBS posts reflect daily life, and then long interval models only catch those topics lasting longer and "ignore" some suddenly happened while soon disappeared hot events. Another problem is to define the appropriate span of time. DTM assumes that the number of topics is fixed and the topics extend along the span of time. The foci of on-line

[4] The DTM code package can be downloaded from http://www.cs.princeton.edu/~blei/.

discussions are often updated quickly and many last only several days, such as topics on the holidays. Thus, long time span may not be appropriate. Moreover, consider the model efficiency, if with too many time stamps, posterior inference will require massive amounts of time and memory.

In our research, we choose 13-month period, Dec, 2011 to Dec, 2012 respectively as the data source to train the model respectively with 60 topics and 1-day as the time slice.

4.1 Words Variety Exploration

At the corpus level, each topic is now a sequence of distributions over words according to the posterior inference $\beta_{k,t}$. In practice, different terms may be used in different periods. Then the distributions over words reflect the central viewpoints shift along the on-line discussion.

Take the model results of December of 2011 for example. Topics with words about "school bus" and "environment" are illustrated for detailed analysis. Figure 1 shows the trends of both hot and cold words for topic "school bus", while Figure 2 is for topic "environment". The y-axis is the probability of words under the respective topic. Hot words are picked according to the biggest positive difference between the final and initial probability of each word of that month, i.e. $top(\beta_{k,31}-\beta_{k,1})$, and cold words are taken based on the positive value of $top(\beta_{k,1}-\beta_{k,31})$. For better understanding, Tables 3 and 4 give some top words (according to $\beta_{k,t}$) under the respective topic at several time stamps.

By combining the trends analysis (Figure 1 & 2) and the top words course (Table 3 & 4), some conclusions are obtained. In topic "school bus", the terms "accident" and "safety" rise up along the time, whereas the term "school bus" with a big wave reflects the drastic fluctuation of the focus on the school bus accident which happened at Feng Country, Jiangsu Province on December 13.

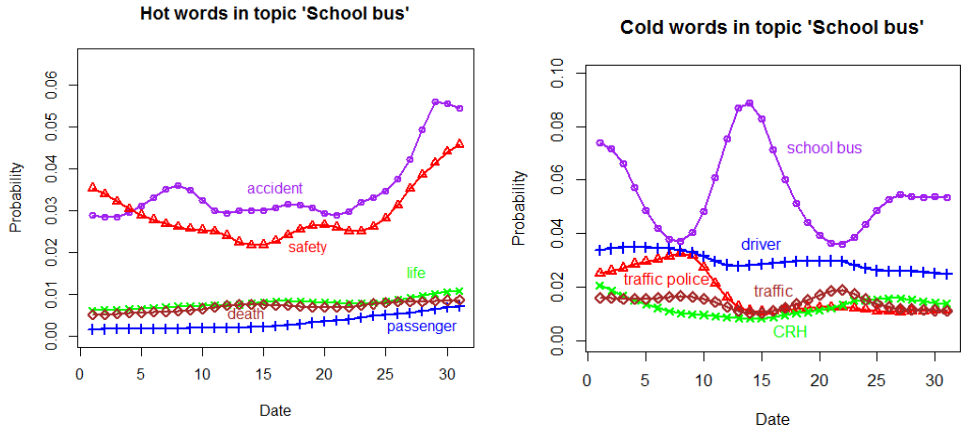


Figure 1 Word trend shifts on the topic "school bus".

Table 3 Top words in some selected time stamps of the topic "school bus"

Time stamp	Top words
Dec. 1	school bus, safety, driver, accident, traffic police, CRH, vehicle, traffic, government sector, blame
Dec. 5	school bus, driver, accident, traffic police, safety, vehicle, traffic, CRH, government sector, blame
Dec. 9	school bus, accident, driver, traffic police, safety, traffic, vehicle, government sector, CRH, problem
Dec. 13	school bus, accident, driver, safety, vehicle, traffic police, Feng Country, student, traffic, government sector
Dec. 17	school bus, accident, driver, safety, vehicle, traffic, traffic police, Feng Country, government sector, student
Dec. 21	school bus, driver, accident, safety, traffic, vehicle, traffic police, CRH, government sector, blame
Dec. 25	school bus, accident, safety, driver, vehicle, CRH, traffic, bus, traffic police, blame
Dec. 29	accident, school bus, safety, driver, vehicle, CRH, blame, traffic, traffic police, life

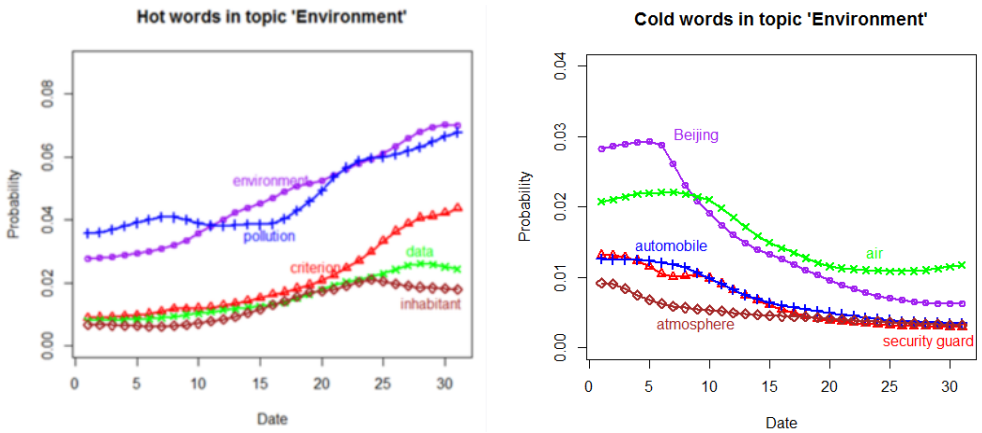


Figure 2 Word trend shifts on the topic "environment".

Table 4 Top words in some selected time stamps of the topic "environment"

Time stamp	Top words
Dec. 1	pollution, Beijing, environment, city, air, environmental protection, security, seriousness
Dec. 5	pollution, Beijing, environment, air, city, environmental protection, seriousness, air quality
Dec. 9	pollution, environment, air, Beijing, city, environmental protection, air quality, seriousness
Dec. 13	environment, pollution, city, air, environmental protection, Beijing, standard, seriousness
Dec. 17	environment, pollution, city, vaccine, standard, environmental protection, inhabitant, data
Dec. 21	Environment, pollution, standard, city, environmental protection, data, inhabitant, vaccine
Dec. 25	environment, pollution, standard, city, data, inhabitant, environmental protection, expert, countrywide
Dec. 29	environment, pollution, standard, city, data, inhabitant, environmental protection, countrywide, air

In topic "environment", the curves about terms "environment", "pollution" and "standard" move up and terms "Beijing" and "air" cool down, which reflect the transition of the topics from Beijing's air quality to countrywide air quality, from discussions of serious environment pollution to the establishment of the air quality measures.

Next we go further to analyze the topic variability in the time series at the document level.

4.2 Topics Evolution

According to the posterior inference of the topics distribution $\theta_{d,t}$, the mixture of topics of each post is obtained. To grasp the change of the strength of each topic, we define an *average* $\theta_{k,t}$ to denote the average strength of topic k at time t. The volatility of $\bar{\theta}_{k,t}$ reflects the changing of public foci. The *average* $\bar{\theta}_{k,t}$ is calculated by following steps:

For all $t = 1,\ldots,T$ (T is the time slice), repeat:

Step 1: sum up the number of posts M_t;

Step 2: at time t, the strength of topic k is calculated as $\bar{\theta}_{k,t} = \frac{1}{M_t}\sum_d \theta_{d,t}$.

Figure 3 is the diagram of the computation.

The calculation results of the *average* $\theta_{k,t}$ of some selected months are visualized with grid graphs in sequence as shown in Figure 4. The y-axis lists 30 topics which are picked by value of $\sum_t \bar{\theta}_{k,t}$ from 60 topics in each month, and ranked downwards. The strength of each topic is presented by interval values.

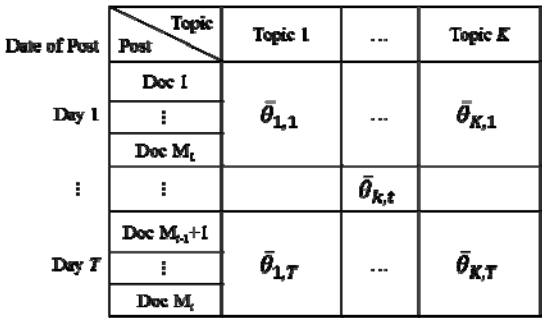


Figure 3 Diagram of the topic strength computation

4.3 Result explanations

At first, we look into the results of some selected months to see how the hot topics fluctuate according to the strength. Obviously some topics are closed related to the real events.

In January of 2012, as shown in Figure 4(b), topics about *Spring Festival Gala* became hot due to the Chinese New Year (January 23). People who cared about buying tickets online during the festival had complaints about the *travel rush* and the *booking website*. Some social events, such as the inappropriate comments given by both Zhang Shaogang and Kong Qingdong, caused public criticisms which lasted for a short time. The topics on *Han Han*[5] and his debate with *Fang Zhouzi* started, and suddenly became hot in the latter of the month.

[5] Han Han, a racing driver and novelist, was in debate with a fraud fighter, Fang Zhouzi, that whether his blogs were written by others.

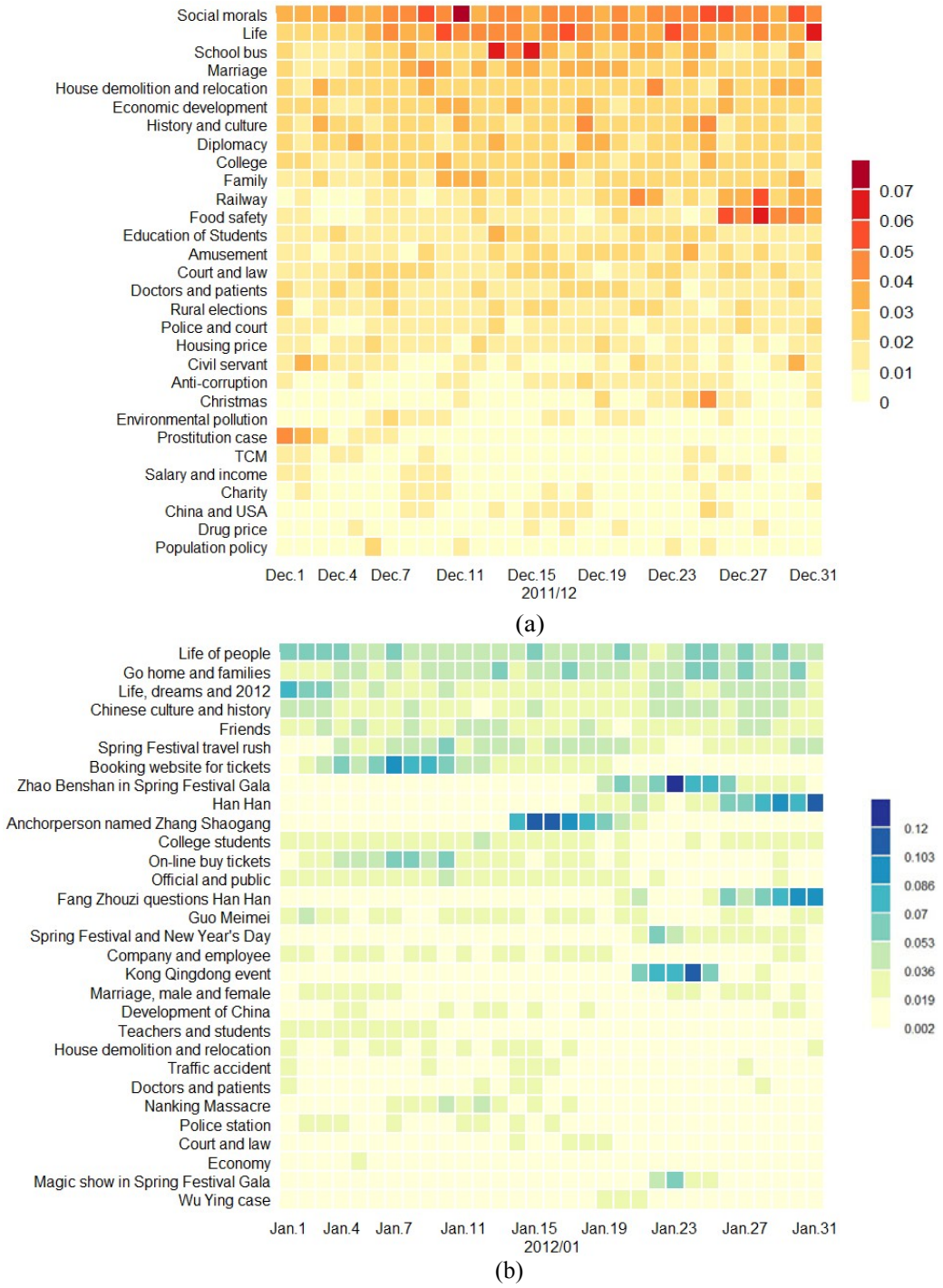
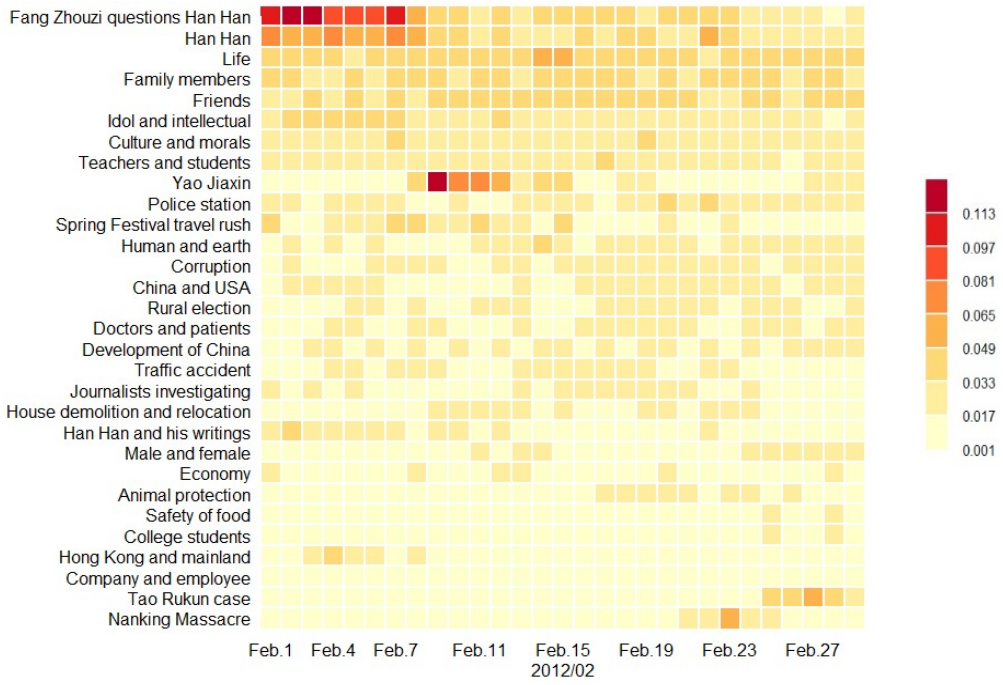
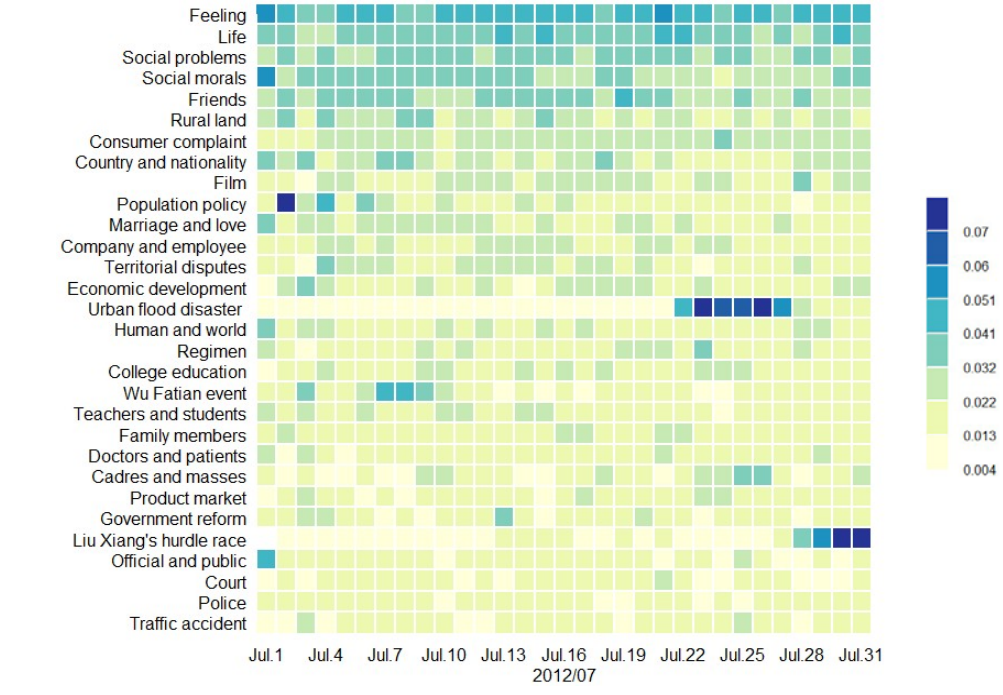


Figure 4 The lateral axis represents the date whereas the vertical axis represents the topics. The grid is the gradation of the strength of topics by $\overline{\theta}_{k,t}$ at time t

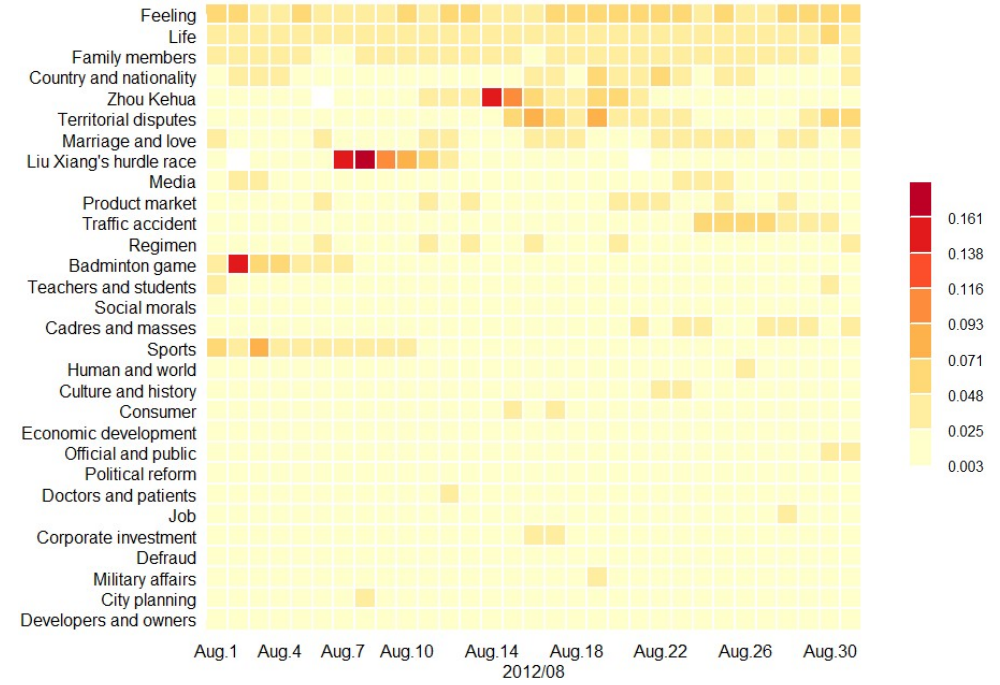


(c)

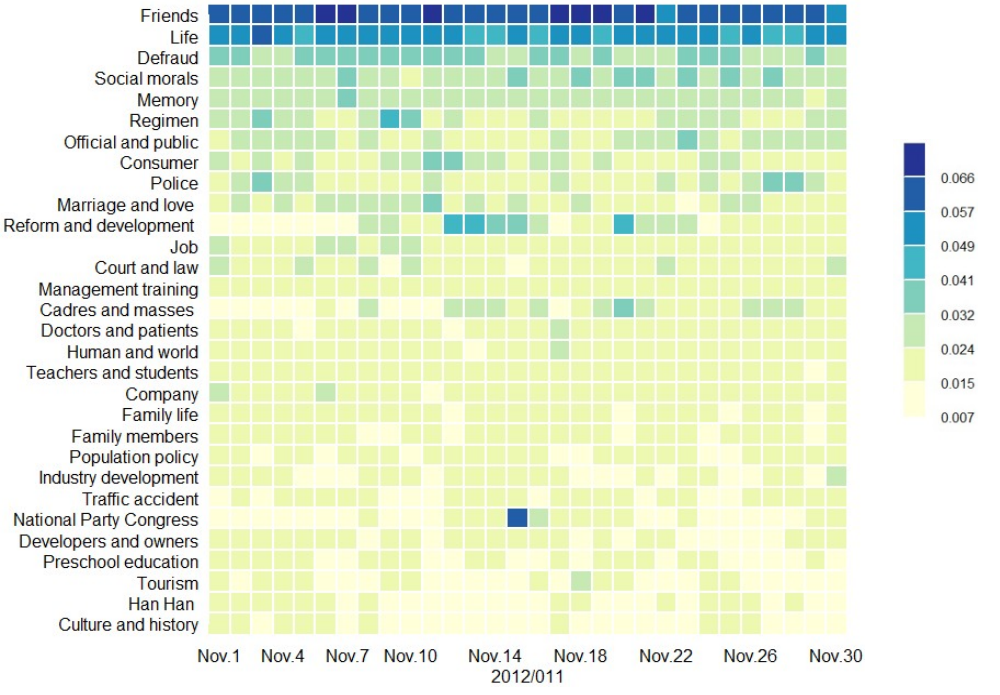


(d)

Figure 4 (*continued*)



(e)



(f)

Figure 4 (*continued*)

In February of 2012, as shown in Figure 4(c), topics about *Han Han and Fang Zhouzi* became quite hot, leading many other related issues discussed by public. The name *Yao Jiaxin* was mentioned again due to the dispute of civil compensation aroused in February 8. Another event, *Tao Rukun* case, a high school student doused petrol and set fire to a girl who rejected him, caused people in an uproar.

In July of 2012, as shown in Figure 4(*d*), the hottest event was *rainstorm disasters* in Beijing. Talks about the hurdle race of Liu Xiang began hot in the end of the month.

In August of 2012, as shown in Figure 4(e), *Zhou Kehua event*, the *Olympic Games* and the performance of *Liu Xiang* are substantially discussed by public.

In November of 2012, as shown in Figure 4(f), the 18th *National Party Congress* is a hot topic, together with *reform development*.

Here we just list and analyze some months results for illustration. By the visualized results, we summary some patterns of topics and their fluctuations on TYZT Board as follows:

1) Common topics which emerge almost every month with different intensities but always keep flat fluctuations, including:

a) daily life topics, such as families, feelings, dreams, marriage, friendships, education, college students, job;

b) social morals topics or related ones, like defraud, consumers complaints;

c) economy and culture topics, such as TCM, history, human and world, economic development;

d) governments related topics such as the legislation, anti-corruption, official and general public, social reform;

e) topics about assertion of rights, such as medical dispute between doctors and patients, house demolition and relocation, rural election, police and court;

f) topics about population policy, environmental pollution, and animal protection.

2) Topics related to specific days, whose fluctuations are quite related to the date and can be predicted to some extent, include:

a) festival topics, for example, Christmas, Spring Festival;

b) topics derive from festivals such as Spring Festival Gala and Spring Festival travel rush.

c) topics related to regularly held large-scale or popular activities, especially sports games, such as the Olympic Games.

3) Topics on social incidents or hot events often erupt and are difficult to be predicted; especially those topics related to people's safety or benefits. They are quite sensitive to either the government decisions or the social events of high societal risk.

a) topics related to people's safety or benefits, such as food safety, travel rush, price rises, school bus, urban storm. People discuss them online in real time.

b) topics on hot criminal cases, such as Tao Rukun case, Zhou Kehua case, Yao Jiaxin case.

c) topics on public figures' comments or behaviors, such as debates of Han Han with Fang Zhouzi, Kong Qingdong event. These

events are not directly related to pubic life but will be discussed totally and widely.

According to the above-mentioned three categories of topics and their fluctuations, some hints for societal management and social stability can be acquired. It is better for local government to pay attention to the first category topics that are about the governance and assertion of rights, because those topics with high societal risk (Zheng, Shi & Li, 2009) may reveal the phenomena of social instability. It is also better to provide clear and right information and prevent the misinformation toward the third category topics which are about hot events, especially under emergency state. It is quite normal for the 2nd category topics and then normal procedures to deal with predictable situations can be designed in advance.

5. Conclusions

Social media is an open system with multiple, diverse and open minds compared to the traditional media. Being aware of the on-line public opinions or public concerns is essential to societal management in current China. Lots of studies are being undertaken to fast catch the primary topics and the trends of topics especially under big data era. Our study is based on on-line forum data, trying to provide an effective way to on-line public opinions detection.

Topic models are applied to analyze new posts on Tianya forum to discover hot topics and their tendency over time. After observing the topics fluctuations in 13 months, three patterns are summarized by a macroscopic perspective. Also some hints are discussed to deal with those possible relevant problems. Topics with higher societal risk are worth more attention, although these topics are quite common.

Beyond all doubts, DTM offers new ways to browse large and unstructured document collections. However according to current literatures the complexity of topic models seems limit their application in China. Some dynamic topic analyses are based on simpler LDA modeling, and the amount of posts is much less (Chu, 2010; Shi & Zhang, 2012). This paper presents a practical attempt to explore DTM to 13-month BBS posts. Theoretically, it is an extension of the model application to Chinese BBS. It is also an attempt to help to deal with social problems using computing ways by statistic models.

Dynamic topic model also has some disadvantages. One is that the number of topics is fixed and the disappearance, expansion and shift of the topics are ignored. Another is that the inference complexity grows quickly as time granularity increases. Then the time granularity cannot be too fine. Besides, there is no objective standard to measure the validation of the model results.

Unlike the studies on on-line behaviors, such as posting and browsing behavior analysis (Cui, He & Liu, 2013), or hits and replies statistics (Zhao & Tang, 2013), we concern more on content analysis. Textual analyses are much different from behavior analyses on catching the primary topics. The former gains deeper insight into on-line community's behaviors, attitudes, concerns and culture than the latter from social

science perspective, while a combined approach is preferred.

Lots of works need to be improved. In the future, we will analyze the updated posts which reflect daily on-line concerns on social events more than original posts so as to understand the tendency of hot topics better, even the size of data sets increase evidently. Topics at different BBS boards are also required to be studied for comparisons. Moreover, we may consider geographical factors to see how the local culture affects the hot topics. Obviously there is a long way to explore and implement computing ways to bottom-up emerging unstructured on-line textual data for better understanding public concerns comprehensively for effective societal management.

References

[1] Allan, J., Carbonell, J. G., Doddington, G., et al.(1998). Topic detection and tracking pilot study final report. Proceedings of the Broadcast News Transcription and Understanding Workshop (Sponsored by DARPA), Feb, 1998

[2] Alsumait, L., Barbara, D., Domeniconi, C. (2008). On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08), 3-12, Pisa, December 15-19, 2008, IEEE

[3] Blei, D. M., Lafferty, J. D. (2006). Dynamic topic models. In: W. W. Cohen & A. Moore (eds.), Proceedings of the 23rd International Conference on Machine Learning, 113-120, Pittsburgh, PA, June 25-29, 2006, ACM

[4] Blei, D. M., Lafferty, J. D. (2007). A correlated topic model of Science. Annals of Applied Statistics, 1: 17-35

[5] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022

[6] Cao, B., Shen, D., Sun, J., et al. (2007). Detect and track latent factors with online nonnegative matrix factorization. In: M. M. Veloso (ed.), Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2689–2694, Hyderabad, January 6-12 2007

[7] Chu, K. M. (2010). The Research on Topic Evolution for News based on LDA Model. Master Thesis. Shanghai Jiao Tong University. (In Chinese)

[8] Cui, L. J., He, H., Liu, W. (2013). Research on hot issues and evolutionary trends in network forums. International Journal of u- and e- Service, Science and Technology, 6(2): 89-97

[9] Guha, R., Kumar, R., Sivakumar, D., and Jose, S. (2005). Unweaving a web of documents. In: R. Grossman, R. J. Bayardo & K. P. Bennett (eds.), Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Illinois, August 2005, ACM

[10] Hall, D., Jurafsky, D., Manning, C. D. (2008). Studying the history of ideas using topic models. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 363-371, Hawaii, October 2008, ACL

[11] Hao, X., Hu, Y. (2010). Topic detection and tracking oriented to BBS. Proceedings of

2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, 4: 154-157, Changchun, August 24, 2010, IEEE

[12]He, Q., Chen, B., Pei, J., et al. (2009). Detecting topic evolution in scientific literature: how can citations help? Proceedings of the 18th ACM conference on Information and Knowledge Management, 957-966, Hong Kong, November 2009, ACM

[13]Jin, L. (2008). Chinese outline BBS sphere: what BBS has brought to China. Dissertation, Massachusetts Institute of Technology

[14]Kleinberg, J. (2002) Bursty and hierarchical structure in streams. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 91-101, Alberta, July 2002, ACM

[15]Lu, M., Yao, X., Wei, S. (2008). BBS hot topic mining algorithm based on fuzzy clustering. Journal of Dalian Maritime University, 34 (04): 52-58. (In Chinese)

[16]Meng, C., Zhang, M., Guo, W. (2012). E volution of Movie Topics Over Time. UR L: http://cs229.stanford.edu/proj2012/MengZhangGuo-EvolutionofMovieTopicsOverTime.pdf. Cited April 1, 2014

[17]Nallapati, R. M., Ditmore, S., Lafferty, J. D., et al. (2007). Multiscale topic tomography. In: P. Berkhin, R. Caruana, & X. D. Wu (eds.), Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 520-529, California, USA, August 12-15, 2007, ACM

[18]Rittel, H. W. J., Webber, M. M. (1973). Dilemmas in a general theory of planning. Policy Sciences, 4(2): 155-169

[19]Shi, D. W., Zhang, H. (2012). LDA Model-based BBS topic evolution. Industrial Control Computer, 25(05): 82-84.(In Chinese)

[20]Song, X., Lin, C. Y., Tseng, B. L., et al. (2005). Modeling and predicting personal information dissemination behavior. In: R. Grossman, R. J. Bayardo, & K. P. Bennett (eds.), Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 479-488, Illinois, August 21-24, 2005, ACM

[21]Tang, G., Chen, H. (2010). Text clustering method based on BBS hot topics discovery. Computer Engineering, 7: 31. (In Chinese)

[22]Wang, C., Blei, D., Heckerman, D. (2008). Continuous time dynamic topic models. In: D. A. McAllester & P. Myllym (eds.), Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, 579-586, Helsinki, July 9-12, 2008, AUAI Press

[23]Wang, X., McCallum, A. (2006). Topics over time: a Non-Markov continuous-time model of topical trends. In: T. Eliassi-Rad, L. H. Ungar, M. Craven, et al. (eds.), Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 424-433, Philadelphia, August 20-23, 2006, ACM

[24]Wayne, C. L. (2000) Multilingual topic detection and tracking: successful research enabled by corpora and evaluation. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC), Greece, May 31 -June 2,

2000, European Language Resources Association

[25] Wu, J. J., Sun, H. Y., Tan, Y. (2013). Social media research: a review. Journal of Systems Science and Systems Engineering. 22(3): 257-282

[26] Yang, Y., Pierce, T., Carbonell, J. (1998). A study of retrospective and on-line event detection. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 28-36, 1998, ACM

[27] You, L., Du, Y., Ge, J., et al. (2005). BBS based hot topic retrieval using back-propagation neural network. In: K. Y. Su, J. Tsujii, J. H. Lee, ea al. (eds.), Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP 2004), LNCS 3248:139-148, Hainan Island, March 22-24, 2004, Revised Selected Papers, Springer

[28] Zhang, J., Li, F. (2012). LDA topic evolution based on global and local modeling. Automation Technique, Computer Technology. 46 (11): 1753-1758

[29] Zhang, Z. D., Tang, X. J. (2011). A preliminary study of web mining for Tianya forum. Proceedings of the 11th Youth Conference of Systems Science and Management Science and 7th Conference of Logistic Systems Technology. Wuhan: Wuhan University of Science and Engineering Press, 199-204. (In Chinese)

[30] Zhao, Y. L., Tang, X. J. (2013). A preliminary research of pattern of users' behavior based on Tianya forum. In: S Y Wang, Nakamori Y and W L Jin, (eds). Proceedings of the 14th International Symposium on Knowledge and Systems Sciences, 171-179, Ningbo, October 25-27, 2013, JAIST Press

[31] Zheng, R., Shi, K., Li, S. (2009). The influence factors and mechanism of societal risk. In: J. Zhou (Ed.): First International Conference on Complex Sciences, Complex 2009, Part II, LNICST 5, 2266-2275, Shanghai, February 23-25, 2009, Springer

Lina Cao is a doctoral student in Academy of Mathematics and Systems Science, Chinese Academy of Sciences. Her research interests include knowledge science, systems science, text mining. She just finished her oral defence for her dissertation in May of 2014.

Xijin Tang is a full professor in the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. She received her BEng (1989) on computer science and engineering from Zhejiang University, MEng (1992) on management science and engineering from University of Science and Technology of China and PhD (1995) from CAS Institute of Systems Science. During her early system research and practice, she developed several decision support systems for water resources management, weapon system evaluation, e-commerce evaluation, etc. Her recent interests are meta-synthesis and advanced modeling, opinion dynamics, knowledge creation and creativity support systems. She co-authored and published two influential books on meta-synthesis system approach and an oriental systems approach in Chinese. She is the secretary general of International Society for Knowledge and

Systems Sciences. She was one of 99 who won the 10th National Award for Youth in Science and Technology in China in 2007.