KSS'2013 NINGBO

Proceedings

Knowledge Creation Towards Emergency Management

Shouyang Wang, Yoshiteru Nakamori and Weiliang Jin (eds.)

JAIST Press

ISBN: 978-4-903092-36-2

Analysis of Dynamic Topics Evolution Based on Tianya Club

Lina Cao **Xijin Tang**
Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing, 100190, China
{caolina, xjtang}@amss.ac.cn

Abstract

On-line public opinion is an important part of social public opinion in today's society. How to grasp the on-line public opinion and detect the hot topics are wicked problems and also new challenges for governance. In our research, in order to structure such wicked problems, Dynamic Topics Models (DTM) and Latent Dirichlet Allocation (LDA) are applied to explore the changing topics of new posts collected from Tianya Zatan Board of Tianya Club. Those online data reflect the community opinions toward social problems.

Keywords: topic models, Dynamic Topic Model, Latent Dirichlet Allocation, topics evolution, Tianya Club

1 Introduction

Social media have become popular platforms for common people to express their opinions on hot social events. Public opinions may drive big changes toward public life, and exert a broad and far-reaching influence on community, traditional media and government decisions. It is of great value to monitor the evolution of hot topics, and to make timely interventions for better management.

The on-line discussions on topics are usually updated, sometimes generating various consequences along the time. For example, last year, "Guo Meimei" event occurred in China. The show-off of luxurious lifestyle by a 20-year-old girl named Guo Meimei and her relations with Red Cross incurred the trust crisis of Chinese Red Cross, and exerted negative effects toward the donation after Lushan earthquake in April of 2013. 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), which public opinions change from the case to the institution, governance, or society [1]. Also the hotness of events are changing. Many things happened suddenly incur the discussions and cooled down slowly as time goes on.

In our research, we try to detect and mine dynamic topics over textual data on Tianya Zatan board [2], the 2nd largest board within Chinese Tianya Club. Comparing to other social media, the on-line forum provides a more interactive conversation about a particular topic than blog. Meanwhile it enables people to post more detailed discussions than microblog, such as Sina Weibo.

In this paper, topic models are used to study the fluctuation of topics in a period and the details of words evolution in the topics are analyzed. The paper is organized as follows. Firstly, the related work is reviewed and the models, Latent Dirichlet Allocation (LDA) and dynamic topic model (DTM), are briefly introduced. Secondly, the data source is presented, and the results of the data analysis are given. Last, conclusions are given.

2 Topic Modeling

2.1 Review

Considering time information for the task of identifying and tracking topics in time-stamped text data is the focus of recent studies in machine learning field [3-5]. Statistical models, topic models, have been applied to solve this task. Topic modeling has become a powerful tool for "extracting surprisingly interpretable and useful structure without any explicit 'understanding' of the language by computer" [6].

According to discrete or continuous time, topic models can be generally divided into three

categories. The first is basic topic model, i.e. Latent Dirichlet Allocation (LDA) model, which treats documents as *bags of words* generated by one or more topics [7]. This model is applied to the whole documents to induce topics sets and then classify the subsets according to the time of documents [8]. The second kind, such as dynamic topic models (DTM) [9] and online LDA (OLDA) model [10], marks documents according to the discrete time before the generative process. The third kind includes Topics over Time (TOT) model which captures both word co-occurrences and localization in continuous time [11] and the continuous dynamic topic model (cDTM) which replaces the discrete state space model of the DTM with its continuous generalization, Brownian motion [12].

Those methods are being successfully applied to explore and predict the underlying structure of various data, such as research papers [10-12], newswire articles [11], personal emails [12] and movie synopsis [13]. In this paper, LDA model and DTM are used.

2.2 Latent Dirichlet Allocation

The basic idea of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The terms are defined by the following: a corpus D is the collection of M documents denoted by $D = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_M}\}$; a document is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_N)$ where w_n is the nth word in the sequence; a word is the basic unit of dicrete data. LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus D [7]:

1. Choose $N \sim$ Poisson(ξ).
2. Choose $\theta \sim$ Dir(α).
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim$ Multinomial(θ).
   (b) Choose a word w_n from $p(w_n|z_n;\beta)$, a multinomial probability conditioned on the topic z_n.

2.3 Dynamic Topic Model

Dynamic topic models (DTM), an extension of LDA, suppose that the data is divided by time slice. The following terminologies and notations can describe the data, latent variables and parameters in the DTM.

- Per-document topics. Each document is a mixture of topics and these different structures cause 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 $\beta_{t,k}$ 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.
- 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 in 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 this model, the multinomial distributions α_t and $\beta_{t,k}$ are generated from α_{t-1} and $\beta_{t-1,k}$, respectively. The generative process for slice t of a sequential corpus is as follows [9]:

1. draw topics $\beta_t|\beta_{t-1} \sim N(\beta_{t-1}, \sigma^2 I)$.
2. draw $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$.
3. for each document:
   (a) draw $\eta \sim N(\alpha_t, a^2 I)$.
   (b) for each word:
      i. draw $Z \sim Mult(\pi(\eta))$.
      ii. draw $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$.

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

$$\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})},\ \theta = \pi(\eta) = \frac{\exp\{\eta\}}{\sum_i \exp\{\eta_i\}}.$$

In DTM, how to learn the parameters according to the only observable $W_{t,d,n}$ constitutes an inference problem. Blei and Lafferty consider that applying Gibbs sampling to do inference in this model is more difficult than in static models, due to the nonconjugacy of the Gaussian and multinomial distributions. They propose the use of variational methods, in particular, the variational Kalman filtering and the variational wavelet regression.

In the research, DTM and LDA are applied to

analysis the data with time-stamps.

3 Topics Detection

Tianya Club has been once the biggest Internet forum in China with approximately 91% visitors comes from China mainland. It has become a comprehensive virtual community that combines online forums, social media and blogging. Among a variety of boards, Tianya Zatan board, the 2nd largest board within the Club, is a specific board of discussion daily lives, social unfair, corruption, phenomena of society, etc. Thus, social events always lead to hot posts with large number of clicks and replies in that board.
In order to study the social stability in current China society, we start to collect data from Tianya Zatan board and several other relevant boards since October of 2010. Now there are about 2,000 new posts published and 4,000 plus posts updated every day. The accumulative data are about 3 million posts. In this paper, we test topic modeling to the posts of this board.

3.1 Data Processing

Before estimating the parameter of the model, data processing is needed. The basic data are new posts (the first post of each new thread) from December of 2011 to March of 2012. Firstly, the posts which have urls but no contents are removed. Secondly, the articles are segmented to words using ICTCLAS[1]. Here Baidu hot search news words are adopted as reserved words [14]. After segmentation, we select nouns and gerunds for analysis. Then, terms that occur fewer than 50 times in corpus and in fewer than 10 posts are removed. The 4-month data are cleaned respectively as the data source. Tab.1 listed the number of posts, corpus and dictionary in each month.

Table 1. The statistics of data sets

Time span / Data statistics	Dec. 2011	Jan. 2012	Feb. 2012	Mar. 2012
New posts #	12,155	12,032	20,124	37,549
Corpus #	14 M	12 M	23 M	45 M
Dictionary #	4,541	3,973	6,091	9,516

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

3.2 DTM Applications

In DTM, 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 confusion task. In the research, we attempt a series of experiments with 1-day interval, 3-day interval and 7-day interval running through DTM package[2] and the results are analyzed using R. It seems that it is not suitable to train our data with too long interval. Unlike the research articles that the new papers definitely are based on the existing research with good consistency, the BBS posts reflect daily life, then long time interval models only catch those topics last longer and "ignore" some suddenly happened soon disappeared hot topics. Another problem is to define the appropriate span of time. DTM assumes that the number of topics is fixed and the topics extend the span of time. On the Web, topics are updated quickly and many last only several days, such as topics on the festival. Thus, long time span may not be appropriate.

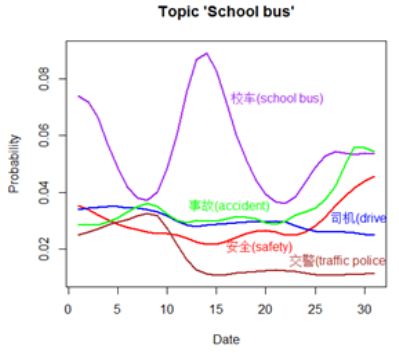
In our research, we choose 4-month period, Dec, 2011 to Mar, 2012 respectively as the data source to train the model respectively with 60 topics and 1-day as the time slice.

Words Variety Exploration. At the topic level, each topic is now a sequence of distributions over words according to the posterior inference. In practice, different terms may be used in different periods when people discussing. Thus, the distributions over words reflect the changing of the key point of the discussion. We select topics with words about "school bus" and "environment" for detail analysis. Figure 1 and 2 show the varying top several words distributions in this topic through different perspectives. Figure 1a and 2a show the trends of words along the whole December. The y-axis is the probability distribution of words under the respective topic. Figure 1b and 2b show top 6 words within the topics in several time stamps.

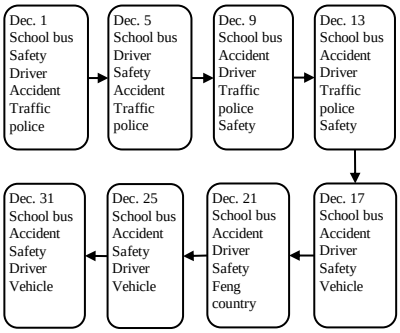
We go further to analyze the topic variability

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

in the time series from a level of corpus.

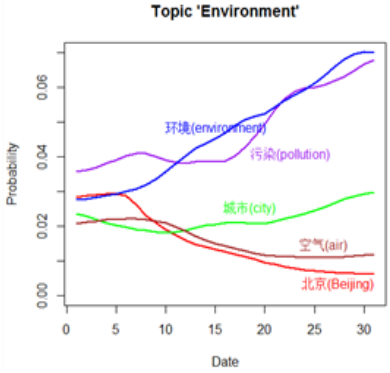


(1a)

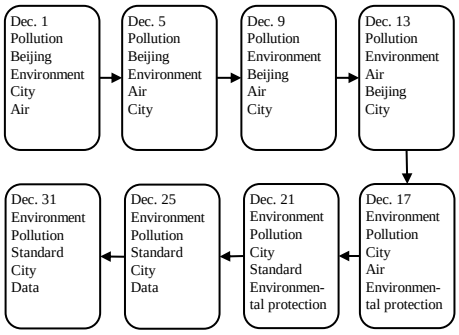


(1b)

Figure 1. Results of word trend analysis on topic "school bus".



(2a)



(2b)

Figure 2. Results of word trend analysis on topic "environment".

Topics Evolution. According to the posterior inference of the topics distribution, the mixture of topics in each post is obtained. Let $\theta_{d,t}$ denotes the topic distribution for post *d* in time *t*. We calculate an *average* θ_t (denotes as $\overline{\theta_t}$) for each post on each day to depict the average hot degree of topics per-day. The volatility of $\overline{\theta_t}$ reflects the changing of public foci on topics. The *average* θ_t is calculated by following steps:

Step 1: Sort all posts according the time stamp and count the number of posts M_t in each time slice;

Step 2: For all $t = 1,\ldots,T$ (T is the time slice), repeat:

*Step 2.*1: at time *t*, sum up the proportions of posts for each topic, denotes that $\theta_t = \sum_d \theta_{d,t}$;

Step 2.2: for each topic, θ_t divided by the number of posts, M_t, denotes that mean $\overline{\theta_t} = \frac{1}{M_t}\sum_d \theta_{d,t}$.

Figure 3 is the diagram of the computation.

	Post	Topic 1	…	TopicN
Day 1	Doc 1		$\overline{\theta_1}$	
	……			
	Doc M_1			
	… …			
Day *t*	Doc M_{t-1}+1		$\overline{\theta_t}$	
	……			
	Doc M_t			

Figure 3. Diagram of the computation

The calculation results of the *average* θ_t of four months are drawn in Figure 3 in sequence. In each month, the y-axis is 30 topics which are picked with high value of θ_t and ordered by the value.

Result explanations. We analyze the results which are calculated by the algorithm month by month. According to the series of experiments, it can be directly perceived some regular patterns of the fluctuation of topics from Figure 4. First of all, the features of the public opinions on Tianya Zatan Board are the following:

1) topics about *life*, *family*, *friends* and *society morality* are highly and continuously concerned by people. Obviously those common people concern their own daily living mostly;

2) education, festivals, college student, doctor and patient, traffic, economy, culture, official and public etc. remain as general societal topics which are related to people's life;

3) complaints toward power, such as *land demolition*, *the villagers' election*, *court*, *police* and so on, also are main topics posted by people;

4) when the social events happened are related to people's safety or benefits, such as *food safety*, *travel rush*, *price rises*, *school bus*, people discuss real time online; they are quite sensitive to the government decisions and social events which contain societal risk.

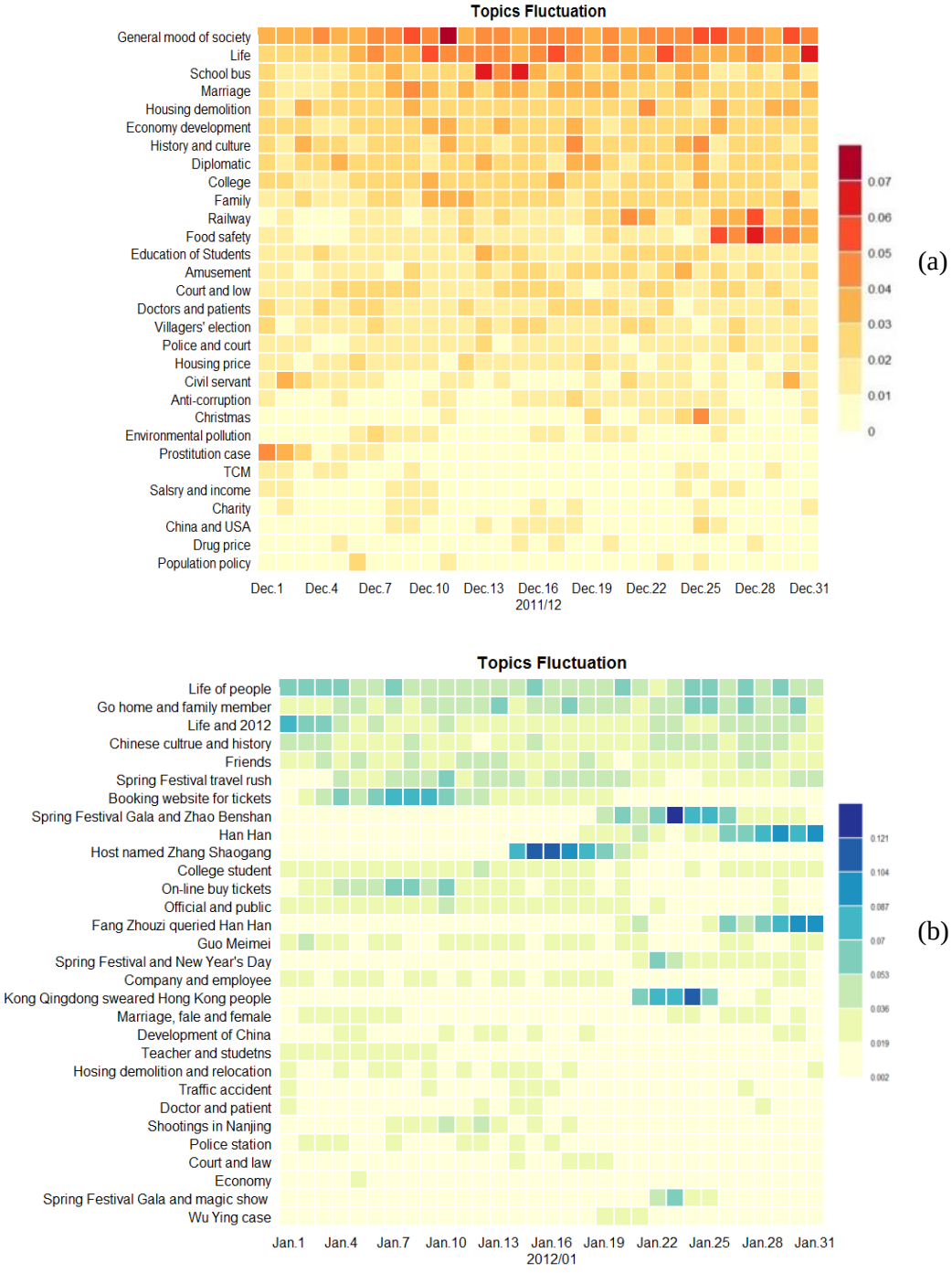


Figure 4. The lateral axis represents the date whereas the vertical axis represents the topics. The cell represents the value of *average* θ_t in each day and its color indicates the interval of the value.

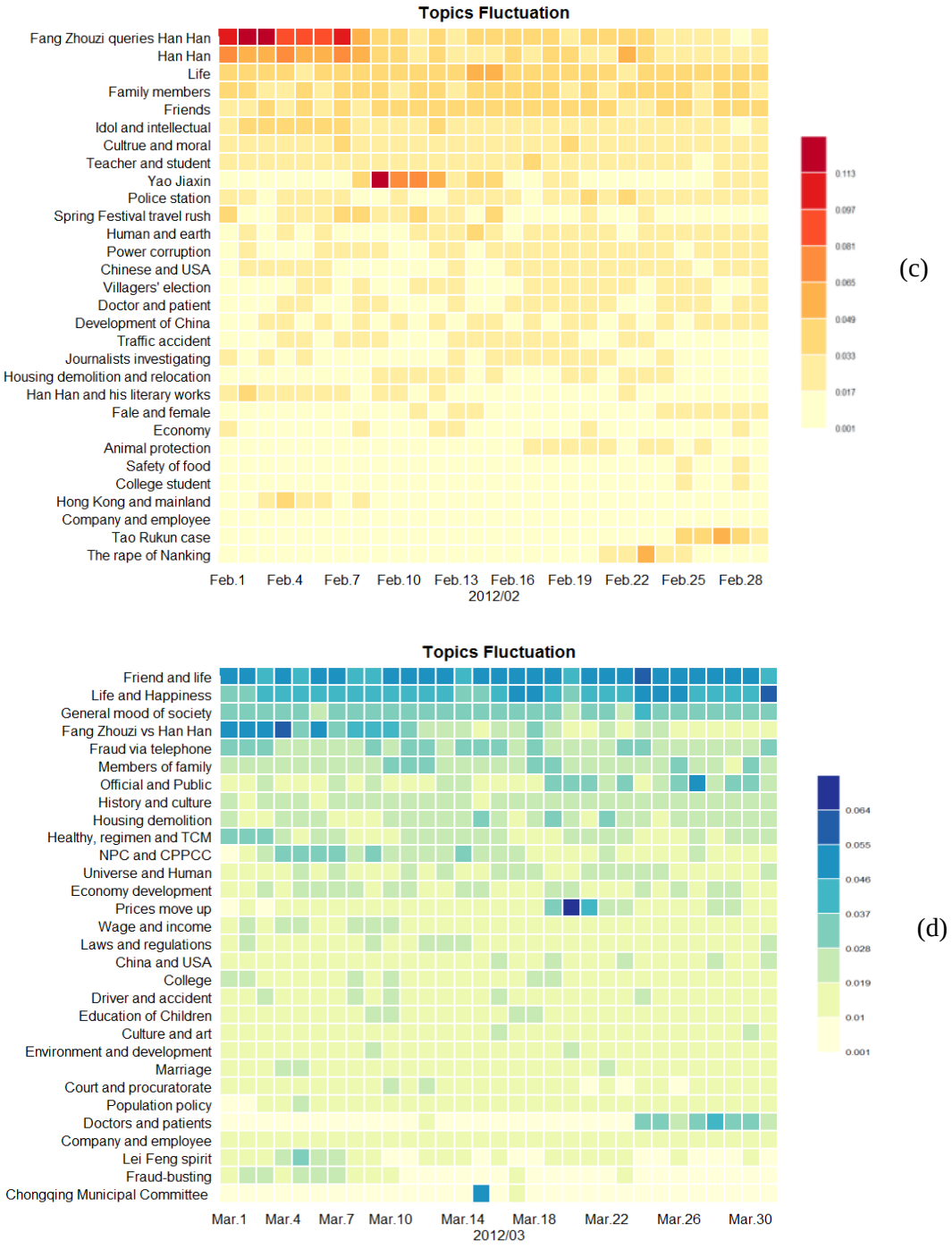


Figure 4. (*continued*)

Besides, some special events happened each month and caused discussion. In December of 2011, as shown in Figure 4 (a), the bad news such as milk poor quality and school bus accident caused the topic *food safety* and *school bus* growth due to their importance. Some topics were related to specific days such as *Christmas* and *railway*. People talked about *railway* because the New Year was coming and many people might go to travel by train, then a small rise of the probability of this topic was around December 21, the first day to allow on-line ticket order.

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). Also,

people who care about the tickets booking 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*[3] (a ricing driver and novelist, was in debate with a fraud fighter, Fang Zhouzi, that whether his blogs were written by others.) and his debate with *Fang Zhouzi* started, and suddenly became hot in the latter part of the month.

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

In March of 2012, as shown in Figure 4 (d), the topic on *Fang and Han* kept hot in early of this month and cooled down slowly in latter part of the month. The topics about political conferences, NPC (National People's Congress) and CPPCC (Chinese People's Political Consultative Conference) hold during March 3-14 drew the concerns of public. Meanwhile, those topics attracted attentions on the political changes in Chongqing and the relationship between *official and public*. The *rise of oil price* issued by Development and Reform Commission in March 19 led to an eruption of discussion, which however it did not last long. Another example of topic related to time was the discussion on *Lei Feng spirit*[4].

3.3 LDA Application

Four models are trained based on the four month data using LDA method. We also calculate the *average* θ_t according to the calculation principle as shown in Figure 3. Here we choose the results of December of 2011 and January of 2012, showed in Figure 5, as examples to contrast the two methods. In this figure, the lateral axis represents the date; the vertical axis represents the topics. The cell represents the value of *average* θ_t which was calculated by the proportions of documents in each day and its color indicate the interval of the value.

Compared to results by DTM, results by LDA show more flat tendency of average degree of hotness and fewer hot events. In general, the topics represented by two methods are similar. Specifically, it is showed in December of 2011 that people are very concerns on the case of prostitution happened in Xi'an at the beginning of the month. In January of 2012, the topics about the Spring Festival were quite hot: *go home and buy tickets,* the *gala* and the *custom* of the festival.

4 Conclusion

In reality, catching both the central topics and the trend of topics from on-line public opinions are wicked problems which are complex, persistent, intractable and above all undefinable [15]. Solutions to wicked problems are not true-or-false, but better or worse. Our study is based on Tianya forum data, trying to provide another access to on-line public opinions perception.

Unlike the researches on on-line behaviors, such as posting and browsing behavior analysis [1], or hits and replies statistics, we concern more on content analysis and try to find the general regularity and characteristics. We consider that hit or reply behaviors represent the responses of onlookers whereas the contents of posts represent the "voices" of common people. For example, many posts focus on the social unfairness; those posts may be with fewer hits than the posts discussing of super stars. But the former has more social influence than latter. Thus, textual analyses are much different from behavior analyses on catching the central topics.

We apply topic models to analyze new posts on forum to discover dynamic topics and their tendency over time. Beyond all doubts, DTM and LDA offer new ways to browse large and unstructured document collections. However, these models also have some disadvantages, such as the low efficiency due to the complexity.

In the future, we will analyze the updated posts which reflect people's discussion on social events so as to understand the tendency of hot topics better.

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

[4] Lei Feng was characterized as a selfless and modest person who was a soldier of China. 5 March is the official "Learn from Lei Feng Day".

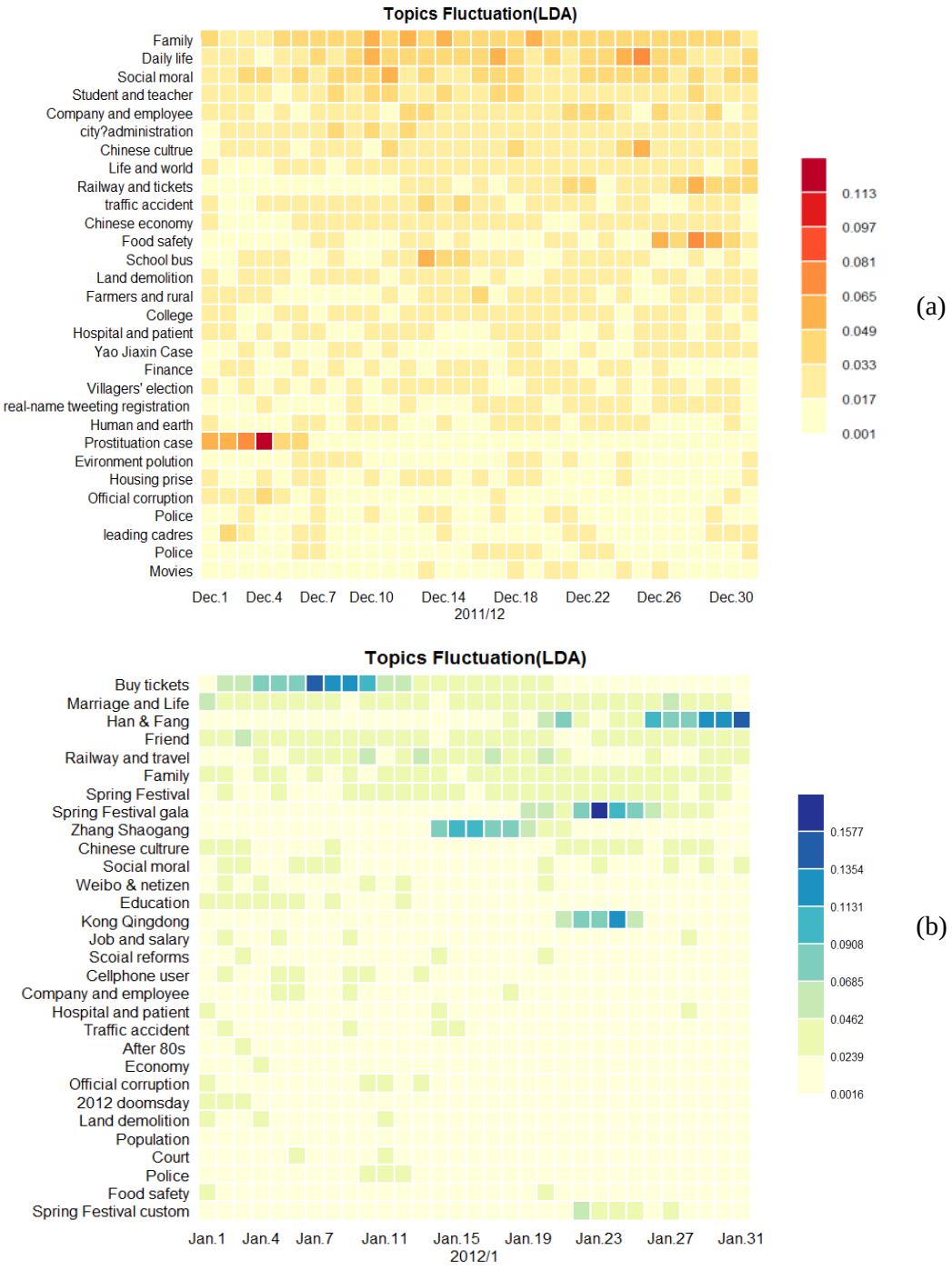


Figure 5. Topics fluctuation shown by LDA results

Acknowledgment

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.

References

[1] Cui, L. J., He, H., Liu, W.: Research on Hot Issues and Evolutionary Trends in Network Forums. International Journal of u- and e-Service, Science and Technology, Vol. 6, No. 2, 89-97. (2013)

[2] Zhang, Z. D, Tang, X. J.: 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 (2011) (In Chinese)

[3] Cao, B., Shen, D., Sun, J., Wang, X., Yang, Q., and Chen, Z.: Detect and Track Latent Factors with Online Nonnegative Matrix Factorization. The 12th International Joint Conference on Artificial Inteligence, pp. 2689–2694 (2007)

[4] Guha, R., Kumar, R., Sivakumar, D., and Jose, S.: Unweaving a Web of Documents. Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, (2005)

[5] Kleinberg, J.: Bursty and hierarchical structure in streams. Proceeding of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, (2002)

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

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

[8] Hall, D., Jurafsky, D., Manning, C. D.: Studying the history of ideas using topic models. Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 363-371 (2008)

[9] Blei, D. M., Lafferty, J. D.: Dynamic Topic Models. In Proceedings of the 23rd International Conference on Machine Learning, (2006)

[10] Alsumait, L., Barbara, D., Domeniconi, C.: On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. ICDM'08. Eighth IEEE International Conference on. IEEE, 3-12 (2008)

[11] Wang, X., McCallum, A.: Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In: KDD'06, USA (2006)

[12] Wang, C., Blei, D., Heckerman, D.: Continuous time dynamic topic models. In Uncertainty in Artificial Intelligence [UAI], (2008)

[13] Meng, C., Zhang, M., Guo, W.: Evolution of Movie Topics Over Time. http://cs229.stanford.edu/projects2012.html. (2012)

[14] Wu, D., Tang, X. J.: Preliminary analysis of Baidu hot words. 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, 478-483 (2011) (In Chinese).

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