Prevailing Trends Detection of Public Opinions Based on Tianya Forum

Lina Cao and Xijin Tang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China (100190)
{caolina,xjtang}@amss.ac.cn

Abstract. How to catch both the central topics and the trend of topics over the on-line discussions are not only of theoretical importance for scientific research, but also of practical importance for social management in current China. In social management perspective, making intervention toward crisis timely and precisely depends on the right image or perception of public opinions toward the crisis. In our research, topic modeling is 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: topics detection, dynamic topic models, Tianya Club.

1 Introduction

In current China, more and more people treat social media as a place to express their opinions about almost everything openly and freely. Toward the social events, besides the traditional media, on-line discussions show fresh, diverse, and evolving opinions which may drive big changes toward public life, and exert a broad and far-reaching influence on community, traditional media and government decisions. It is not only of great value to detect the hazards that people concern [1] and monitor the evolution of hot topics, but also helpful to make reasonable and timely interventions for better management.

In our research, we focus on topics mining over textual data on Tianya Zatan board [2], the 2nd largest board within Chinese Tianya Club. The Forum provides a suitable data source to research the occurrence and evolution of topics about daily lives, social unfair, corruption, phenomena of society, etc.

We use topic models to study the fluctuation of topics in a period and analyze the details of words evolution in the topics in this paper. The paper is organized as follows. Firstly, the related work is reviewed and the dynamic topic model (DTM) is briefly introduced. Then, processing of Tianya Forum posts is addressed, and the results of the data analysis are given. Last, conclusions are given.

2 How to Extract Topics from Textual Opinions

In reality, getting a rough image of a concerned issue is quite important for wicked problem solving, especially the online governance in China. In the past, we did some

H. Yin et al. (Eds.): IDEAL 2013, LNCS 8206, pp. 186–193, 2013.
© Springer-Verlag Berlin Heidelberg 2013

researches using qualitative meta-synthesis technologies, denoted as iView and CorMap, to extract group's opinions and compared the results of different technologies [3]. Those methods are applied to acquire a structure or systemic vision from a group of textual data, draw a scenario using different clustering based on different modeling toward group arguments and extract concepts from clusters of arguments from multi-perspectives [4]. Those studies are exploratory analysis. At those studies, we regard those concepts are topics, central ideas, etc.

Another trend is topic modeling by computer scientists who are on machine learning. Topic modeling has become a powerful tool for "extracting surprisingly interpretable and useful structure without any explicit 'understanding' of the language by computer" [5]. To identify and track dynamic topics from time-stamped textual data, topic models also have been applied. According to discrete or continuous time, topic models can be generally divided into three categories. The first is basic topic model, Latent Dirichlet Allocation (LDA) model, which treats documents as *bags of words* generated by one or more topics [6]. This model is applied to the whole documents to induce topics sets and then classify the subsets according to the time of documents [7]. The second kind, such as Dynamic topic models (DTM) [8] and Online LDA (OLDA) model [9], 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 [10] and the continuous dynamic topic model (cDTM) which replaces the discrete state space model of the DTM with its continuous generalization, Brownian motion [11].

Those models have been successfully applied to explore and predict the underlying structure of various data, such as research papers [9-11], newswire articles [10], personal emails [11], and movie synopsis [12]. Next, we apply DTM to Chinese forum data.

3 Topics Detection from Tianya Club

Tianya Club has been once the biggest Internet forum in China with approximately 91% visitors 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, the 2nd largest board within Tianya Club, is a specific board towards the phenomena of society. Hot events can always been widely and deeply discussed in that board, as illustrated by those posts.

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 Forum Data Processing

We select new posts published during December of 2011 to March of 2012. Data processing excludes posts with empty contents which may be removed due to

censership, includes posts with contents. After segmentation by ICTCLAS[1], Baidu hot search news words are adopted as reserved words [13] and nouns and gerunds are selected 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. Tabel1 lists 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 million	12 million	23 million	45 million
Dictionary #	4,541	3,973	6,091	9,516

3.2 Topic Modeling Experiments

In DTM, time slice is used to divide the data. How to decide the granularity of time is a quite important but confusing task. We attempt a series of experiments with 1-day interval, 3-day interval and 7-day interval running through DTM package[2] and analyze results 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.

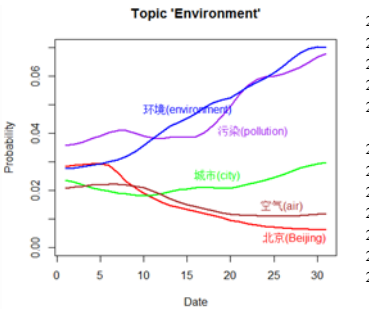
Then, four models with four data sets are trained respectively with 60 topics and 1-day time slice.

3.3 Words Variety

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 a topic with words about "environment" for detail analysis. Fig.1 shows the varying top several words distributions in this topic through different perspectives. Fig.1 (a) shows the trends of words along the whole December. The y-axis is the probability distribution of words under the respective topic. Fig.1 (c) shows top 5 words within the topics in several time stamps. 10 posts which exhibit these topics are selected and their thread titles are listed by time in Fig.1 (b).

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

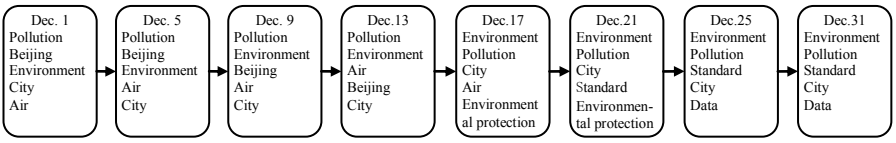
[2] The DTM code package can be downloaded from `http://www.cs.princeton.edu/~blei/`



(a)

2011-12-02 北京 PM2.5 高的快让人窒息了？有木有！
2011-12-06 空气污染与肺癌的相关报道(综合)
2011-12-07 空气环保只能看天意
2011-12-08 世卫城市空气质量排名：北京杭州均列千名之外
2011-12-10 全球 1100 个城市空气质量排名涉及 32 个中国城市，海口排名最高，814 位
2011-12-14 我们的环境好转了吗？
2011-12-15 北京"蓝天"不蓝的重大发现
2011-12-19 中国机动车年排污逾 5000 万吨 成大气污染重要来源
2011-12-20 机动车排放之严重污染恐仍将继续
2011-12-26 "蓝天梦"如此完成：挤出来仅是水分？
2011-12-27 关注 PM2.5 履行企业社会责任
2011-12-31 是轻度污染还是危险？

(b)



(c)

Fig. 1. Results of word trend analysis on topics "environment"

We go further to analyze the topic variability in the time series from a level of corpus.

3.4 Topics Evolution Analysis

According to the posterior inference of the topics distribution, the mixture of topics in each post is obtained. In order to obtain an intuitional sense about the prevailing trends of topics, a statistical variable is designed to depict the evolution.

Algorithm. Let $\theta_{d,t}$ denotes the topic distribution for post *d* in time *t*. In our work, an *average* θ_t (denotes as $\overline{\theta_t}$) is calculated 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: At time *t*, sum up the proportions of posts for each topic, denotes that $\theta_t = \Sigma_d \theta_{d,t}$;

Step 3: For each topic, θ_t divided by the number of posts, M_t, denotes that mean $\overline{\theta_t} = \frac{1}{M_t}\sum_d \theta_{d,t}$;

Step 4: For all $t = 1,\ldots,T$, repeat *Step 2* and *Step 3*.

The calculation results of the *average* θ_t of four months are drawn in Fig.2 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. In December of 2011, topics such as *general mood of society*, *people's life* and *marriage* were always being highly referred. 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 travel by train, then a small rise of probability was around December 21, the first day to allow on-line ticket order.

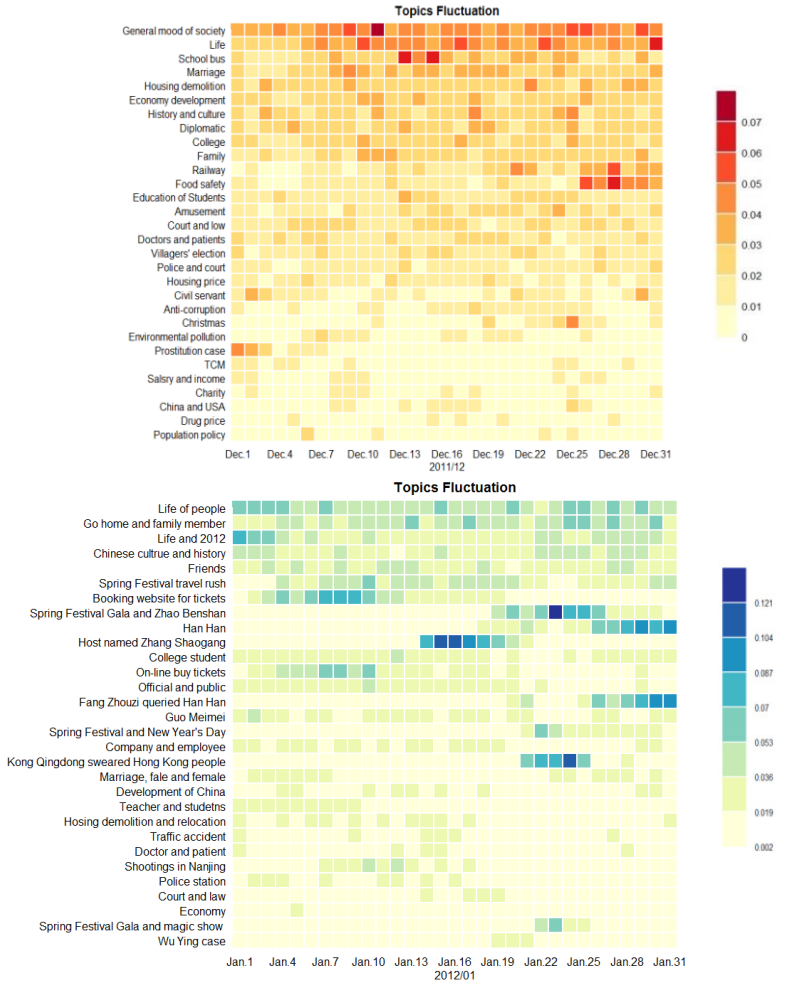


Fig. 2. The horizontal axis represents time whereas the vertical axis represents topics. The cell represents the value of *average* θ_t in each day and its color indicates the interval of the value.

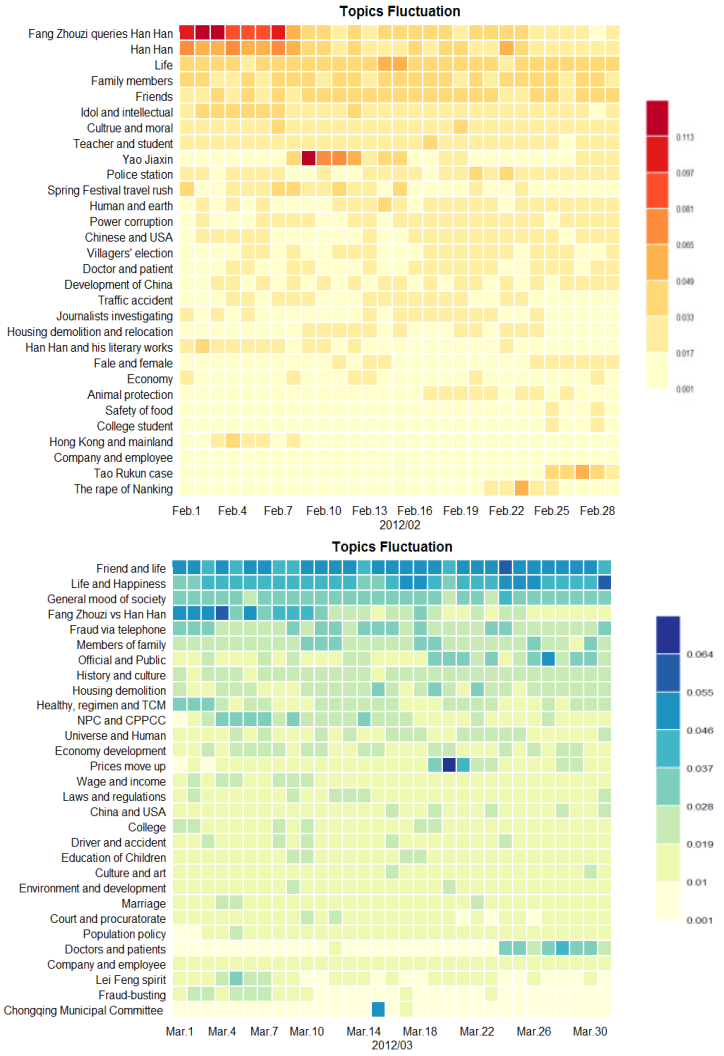


Fig. 2. (*continued*)

In January of 2012, topics about *Spring Festival Gala* became hot due to the lunar new year (January 23). Also, people who care about the tickets ordering 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 last for a short time. The topics on *Han Han*[3] and his debate with *Fang Zhouzi* started, and suddenly became hot in the latter part of the month.

[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.

In February of 2012, 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, 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, but it did not last too long. Another example of topic related to time was the discussion on *Lei Feng spirit*.

In this session, results from a series of experiments illustrate online community have the following characteristics: 1) topics about *life, family, friends* and *society morality* are continuous highly concerns by people. Obviously that common people concern their own lives mostly. *Education, festivals, college student, doctor and patient, traffic, economy, culture, official and public* etc. remain as common societal topics which are related to people's life. 2) 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 contain societal risk.

4 Conclusion

Catching both the central topics and the trend of topics is important for the social management for China. In this paper, we apply dynamic topic models to analyze new posts on forum to discover dynamic topics and their tendency over time. Obviously, DTM offer new ways to browse large and unstructured document collections. Besides the analysis by topic modeling, other methods, such as Self-Organizing Map (SOM) and iView analysis, also can be applied to extract in a meta-synthetic perspective.

Unlike the researches on on-line behavior analysis, such as posting behavior and browsing behavior analysis [14], or hits and replies statistics [15], we concern more on contents and try to find the general regularity and characteristics. The future of our work is that the updated posts which reflect people's discussion on social events will be analyzed to understand the tendency of hot topics better.

Acknowledgements. 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. Zheng, R., Shi, K., Li, S.: The Influence Factors and Mechanism of Societal Risk Perception. In: Zhou, J. (ed.) Complex 2009, Part II. LNICST, vol. 5, pp. 2266–2275. Springer (2009)
2. Zhang, Z.D., Tang, X.J.: A Preliminary Study of Web Mining for Tianya Forum. In: Proceedings of the 11th Youth Conference of Systems Science and Management Science and 7th Conference of Logistic Systems Technology, pp. 199–204. Wuhan University of Science and Engineering Press, Wuhan (2011) (in Chinese)
3. Tang, X.J., Luo, B.: Understanding College Students' Thought Toward Social Events by Qualitative Meta-Synthesis Technologies. International Journal of Organizational and Collective Intelligence 2(4), 15–30 (2011)
4. Tang, X.J.: Qualitative Meta-synthesis Techniques for Analysis of Public Opinions for in-depth Study. In: Zhou, J. (ed.) Complex 2009, Part II. LNICST, vol. 5, pp. 2338–2353. Springer (2009)
5. Blei, D.M., Lafferty, J.D.: A correlated topic model of Science. J. Annals of Applied Statistics 1, 17–35 (2007)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Journal of Machine Learning Research 3, 993–1022 (2003)
7. Hall, D., Jurafsky, D., Manning, C.D.: Studying the history of ideas using topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 363–371. Association for Computational Linguistics (2008)
8. Blei, D.M., Lafferty, J.D.: Dynamic Topic Models. In: Proceedings of the 23rd International Conference on Machine Learning (2006)
9. Alsumait, L., Barbara, D., Domeniconi, C.: On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: Eighth IEEE International Conference on ICDM 2008, pp. 3–12. IEEE (2008)
10. Wang, X., McCallum, A.: Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In: KDD 2006, USA (2006)
11. Wang, C., Blei, D., Heckerman, D.: Continuous time dynamic topic models. In: Uncertainty in Artificial Intelligence, UAI (2008)
12. Meng, C., Zhang, M., Guo, W.: Evolution of Movie Topics Over Time (2012), http://cs229.stanford.edu/projects2012.html
13. Wu, D., Tang, X.J.: Preliminary analysis of Baidu hot words. In: Proceedings of the 11th Youth Conference of Systems Science and Management Science, pp. 478–483. Wuhan University of Science and Engineering Press, Wuhan (2011) (in Chinese)
14. 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 6(2), 89–97 (2013)
15. Chen, X., Li, J., Li, S., Wang, Y.: Hierarchical Activeness State Evaluation Model for BBS Network Community. In: 7th International ICST Conference on Communications and Networking in China (CHINACOM), pp. 206–211. IEEE (2012)