Comput Syst Sci & Eng (2018) 4: 287–291

International Journal of
Computer Systems
Science & Engineering

Topic evolution analysis in social networking services: Taking Sina Weibo as an example

Yuhui Wang

Department of Electronics and Information Engineering Huazhong University of Science and Technology Wuhan, China. E-mail: 18601240913@126.com

Event-related topics in social networking services are always the epitome of heated society issues, therefore determining the significance of analyzing its evolution patterns. In this paper, we present a comprehensive survey on the tweets about "ransomware" in Sina Weibo, a famous social networking service similar to twitter in China. The keyword corresponds to a global ransomware attack in May 2017, on which our example event-related topics are based. We collect text data from sina Weibo and vectorize each tweets, before using a dynamic topic model to discover the event-related topics. The results of the topic model are explainable enough and help us to understand the evolution of those topics more thoroughly

Keywords: Dynamic topic, social networking service, evolution analysis

1. INTRODUCTION

With the social networking services thriving [1, 2], more and more people are willing to share their views and comments, generating as a result are a great many we-media [3] which may largely impact public opinions to some extent. Popular issues and comments in social media always largely reflect dynamic public opinions which will probably have positive or negative impact on society especially in some big events. So far, many researches [4, 5] have been done to investigate public views through social networking services.

To explore how event-related topics evolve in social network services, we focus on the global ransomware attack event in May 2017 and select relative tweets from sina Weibo as the survey target. As is known, there was a ransomware attack spreading through the Internet in an unprecedentedly large scale. The event leaded to serious consequences and adverse social impact, and set off heated discussion in sina Weibo as well. Therefore, we attempt to survey the public discussing topics about that event based on the text data that we crawled from sina Weibo.

Following are the main contribution of the paper.

- We conclude the general research hierarchy for the topic evolution analysis.
- We collect a great many tweets about “ransomware" from sina Weibo, using dynamic topic model to deal with them.
- Based on the topics discovered by the dynamic topic model, we analyze how two example topics evolve associated with the specific condition.

The remainder of this paper is organized as follows. In section 2, we review some related work and introduce the framework of research on social networking services. Then we detail the analysis method in section 3. Subsequently, section 4 analyzes the result of that topic model. Finally, a brief conclusion is provided in section 5.

2. RELATED WORK

Given the current situation that dispersed information greatly saturates sina Weibo, information extraction techniques are firstly required to extract semantically well-defined structured data from unstructured text documents [6]. Meanwhile, it should

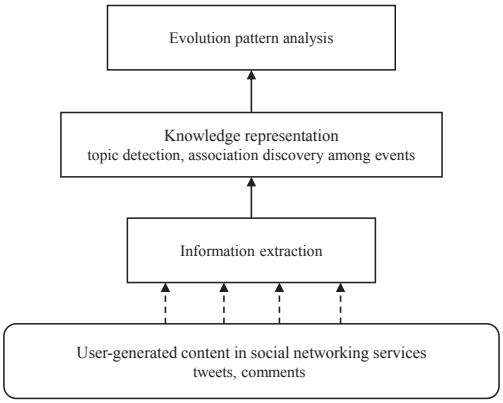


Figure 1 Research hierarchy on social networking services.

be noted that event-related topics beneath those text content involves the knowledge representation from different levels and granularities [7], therefore serving as an assistance to organize the information properly. Furthermore, topic models can be expected to provide evolution pattern analysis with instructive support. Figure 1 shows the research levels based on social networking services, where structured data are transformed from raw user generated content through information extraction and knowledge representation, and evolution pattern analysis are done with the foundation of them.

2.1 Information Extraction

Information surplus, as a common phenomenon in sina Weibo, makes it hard to discover useful information beneath those massive and complex data. When attempting to collect information from sina Weibo, there is a high chance that relevant valuable information remains buried and unobservable, as a result of everyday dramatically increasing content and the restriction of query request. Therefore, certain sampling strategies are required to extract useful information from enormous user generated content. To some extent, the text collection from sina Weibo can be defined as deep-web text collection whose contents are available only via specific search [8].

As a social networking service, sina Weibo exhibit an unstructured view of text data to users, where query-based techniques in pursuit of efficiency and effectiveness should be used to retrieve demanded data and make them structured afterwards [9]. According to [10], a general division exists among currently available query-based techniques, classifying them into two categories, namely bootstrapping approach and statistical learning approach. As for bootstrapping approach, it mainly relies on iteration of retrieval via adopting query operations that start from a “seed" tuple set, a set that is gradually growing thanks to new discoveries in every information extraction step. As for statistical learning approach, documents are labeled as useful and useless in terms of whether they can generate tuples in information extraction step. With the labeled documents, queries to retrieve are learned to distinguish them by training to get expected metric score such as precision and recall.

2.2 Knowledge Representation

With the text data from sina Weibo in structured organization, it is then necessary to discover topics behind them for further research work. Traditionally, topic models based on statistic of text content are widely used to include context information. Latent Dirichlet Allocation (LDA), first proposed in [11], is a generative model that untilizes latent semantic constraint among word, document and topic to estimate the probability where each document belongs to some topics.

Typically, tweets in sina Weibo are rather short, thus standard LDA will fail to handle it. So as to fit with short texts in topic discovery, models specific for tweets have been proposed in [12], where topic distributions of users are taken into consideration and in result more decent metric scores are achieved. Taking temporal relationships into consideration, [13] proposed TM-LDA model via putting emphasis on the topic shift in the data. Describing topic formally is a prerequisite for subsequent analysis, and a number of researchers have been involved in field attempting to find a way to map raw data space to a higher abstract space, where topic can be regarded as an abstract semantic notion compared with raw text data in this case. Some also attempt to solve the encoding problem with topic. For example, Ruben Fernandez Beltran et al. [14] present an approach to encode each sample based on latent topics in Content-Based retrieval.

2.3 Evolution Pattern Analysis

In social networking services, discussions around certain topic always reflect the public opinion which might further have profound influence on society. Therefore, it is quite important to understand how public opinion evolves. So far, there are much research work about the analysis of its evolution pattern in social networking services and the latent meaning behind the user generated contents.

Generally, research efforts are made from empirical perspective and theoretical perspective. From empirical perspective, Liu et al. [15] use LDA to convert short tweets into feature vectors so as to give quantified measurement for political legitimacy of populace, where empirical analysis simply arises from the statistic data. From theoretical perspective, researchers prefer to focus on mathematical models while studying the dynamics in social network. Literature [16] discusses Bayesian and non-Bayesian models to explore the evolution pattern of opinion dynamics. Concerning non-Bayesian models, DeGroot-Friedkin model proposed in [17] tries to predict the evolution of individual social power during the discussion process of several issues.

3. METHOD DESCRIPTION

Our goal is to analyze how the topic evolves based on the text data crawled from sina Weibo, where we choose “Ransomware" as

Table 1 Statistical information of tweets in each period

notation	duration	number
d_1	2017-05-04 ~ 2017-05-16	2,739
d_2	2017-05-16 ~ 2017-05-28	4,628
d_3	2017-05-28 ~ 2017-06-09	1,224
d_4	2017-06-09 ~ 2017-06-21	1,359
d_5	2017-06-21 ~ 2017-07-03	1,847
d_6	2017-07-03 ~ 2017-07-15	237
d_7	2017-07-15 ~ 2017-07-27	51
d_8	2017-07-27 ~ 2017-08-08	285
d_9	2017-08-08 ~ 2017-08-20	352
d_{10}	2017-08-20 ~ 2017-09-01	296
d_{11}	2017-09-01 ~ 2017-09-13	145
d_{12}	2017-09-13 ~ 2017-09-25	247
d_{13}	2017-09-25 ~ 2017-10-07	161
d_{14}	2017-10-07 ~ 2017-10-19	159
d_{15}	2017-10-19 ~ 2017-10-31	1,275

the keyword in this case. We first collect tweets which contain the keyword, then remove some irregular expression in preprocessing step. Subsequently, we run a probabilistic topic model after turning those tweets into bag of words so as to get the most probable terms for each topic [18]. Note that, in order to have a more clear view on the topic evolving process, collected tweets are sliced according to their generated time.

3.1 Data Collection

Sina Weibo contains huge amount of user generated content, however specific content can only be accessed via corresponding search, making it more challenging to collect required data. To overcome that inconvenience, we have realized a targeted crawl tool with the help of Selenium WebDriver in Python which interacts with sina Weibo System automatically by submitting different search criterion. As a result, we have successfully collected 15,185 tweets in specific throughout the duration from May 4th 2017 to October 30th 2017, approximately lasting for half a year.

To discuss the topic evolution pattern, we slice those tweets into 15 parts with the time frequency of 12 days. Table 1 shows the statistic information of tweets in each period. As is shown in Table 1, there is little discussion about ransomware attack in sina Weibo until May 2017 when a terrible epidemic named “WannaCry" outbroke globally. From then on, the number of tweets about “ransomware" increases dramatically during the first two durations, followed by a constant relatively hot discussion in the subsequent three durations. Then the discussion heat drops when the epidemic seems to be controlled in the following nine durations, however, because of the appearance of one variant ransomware named “Bad Rabbit", discussion heat raises again largely yet not as sharp as the scale of first increase in last duration.

For further comprehensive analysis, we attempt to measure the users’ attention degree by calculating the average number of repost, comment and like per tweets for each duration. As is shown in Figure 2, the event of ransomware attack draw public attention when it first broke out on large scale, however, general

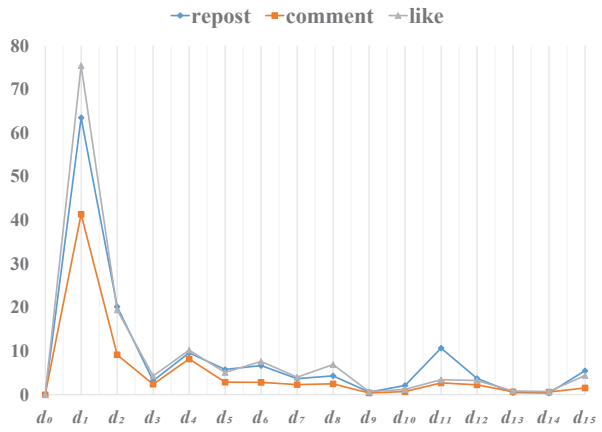


Figure 2 Average number of repost, comment and like per tweets.

attention of users didn’t last long, taking on a prolonged recession afterwards no matter how the discussion continued. Particularly, although the discussion heat was high enough suddenly in last duration, public attention failed to be drawn back.

3.2 Text Preprocessing

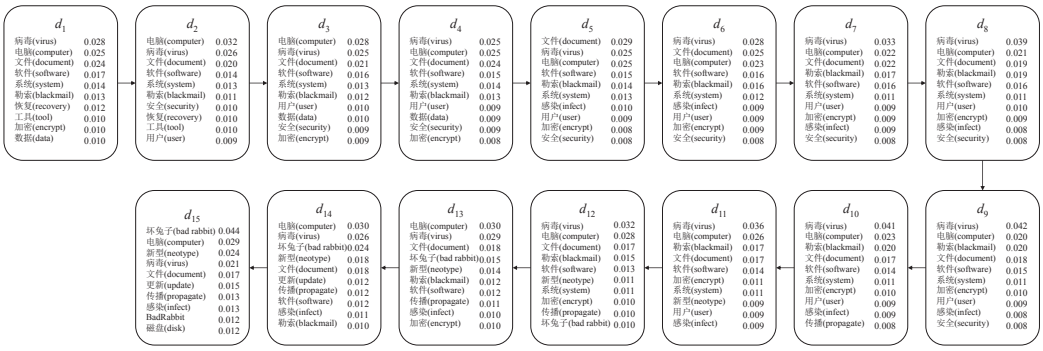
Typically, text preprocessing is the preliminary procedure of topic analysis, since the initial text data crawled from sina Weibo always contain ill formed characters and redundant expression. For instance, the @ sign in sina Weibo represents a mention to the referred user, which is irrelevant with the topic most of the time, and URLs serving as a link to other website in tweets should also be removed. Therefore, we first accordingly attempt to filter those meaningless characters such as @ signs and URLs.

In addition, unlike English text which intrinsically consists of separated words, Chinese text requires word segmentation to split contiguous characters into understandable words. In this case, we apply NLPIR [19], a Chinese lexical analyzer, on filtered text data so as to obtain a list of words for each tweet. Note that in order to get proper word segmentation outcome, we import relevant terms into the user dictionary of NLPIR system beforehand, and remove stop words which commonly appear in those type of text before generating the word lists.

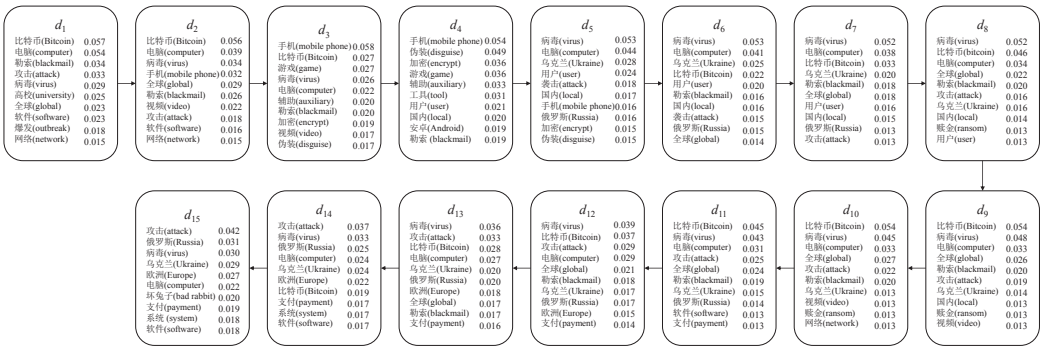
Given the processed word lists, we then subsequently utilize bag-of-words model to represent each tweets and generate a dictionary that orders the appearing words, therefore each tweet can be vectorized according to the frequency of word appearance. With all the tweets consistently vectorized, semantic topics can further be found with the assistance of LDA model.

3.3 Topic Model

As is mentioned in Section 3.1, text data have been sliced into 15 durations for the analysis of topic evolution. Typical topic models, such as LDA model, are capable of dealing with static documents however face dilemma when it comes to dynamic time series documents. Therefore, we adopt a dynamic topic



(a) The evolution of top ten terms in topic 1



(b) The evolution of top ten terms in topic 2

Figure 3 Two examples of 5-topic dynamic model.

model extended from LDA [20] to analyze those dynamic situations.

The dynamic topic model denotes $\beta_{t,k}$ as the V-dimensional vector of natural parameters for topic k in slice t, and the parameters are chained in a state space model that evolves Gaussian noise, following $\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I)$. Unlike LDA, whose topic proportions θ are drawn from a Dirichlet distribution, the dynamic topic model utilize a logistic normal with mean α to express uncertainty over proportions, whose sequential structure between models can be denoted as $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I)$. Using graphical model, the dynamic topic model is gradually simplified to independent topic models as time dynamics, where kth topic at slice $t-1$ smoothly evolves to the topic at slice t.

Particularly, in our case tweets have been divided into 15 time slices with a duration of 12 days, and we set $K = 5$, where K is denoted as the total number of topic. In the next section, we will attempt to analyze how corresponding topic evolves from former slice to succeeding slice.

4. ENVOLUTION ANALYSIS

As is illustrated in Section 3.1, we analyze 15,185 tweets from sina Weibo, all of which are generated between May 4th 2017 and October 30th 2017. The total size of dictionary is 16,571, after removing a great many of stop words that appear too frequently or make little difference in context. Furthermore, we use dynamic topic model with 5 topics to discover the topic of each tweets.

Figure 3 exhibits the evolution process of two resulting topics, showing top ten terms for each topic along with the probability. The dynamic topic model captures discussion topics from different perspectives, where topic 1 emphasize data security while topic 2 focus more on the influencing zone according to Figure 3.

As for topic 1, relative discussions in sina Weibo are more likely to involve data recovery when the ransomware attack outbroke in the first two durations. It is understandable that many normal users attempted to explore emergency recovery methods in social networking services after their important documents were encrypted maliciously, especially as the time when thousands of senior students failed to submit their theses during graduation season. However, after the first two durations, the discussions about data recovery faded and the topic evolution became nearly smooth and steady during duration 3 to duration 10, with less normal users seeking professional assistance and more we-media broadcast relative news. That process lasted until an attention transition took place, when successive discussions put more emphasis on a neotype virus and its potential information security hazard.

As for topic 2, duration 1 and 2 mainly discussed about the general baneful influence, with words like “global" and “network" included. Note that “university" was included in the topic of first duration since most initial victims are university students. Subsequently, “mobile phone", “auxiliary", “tools" and

“game" took up the topic, with the initial virus serving as an analogy, since at that time ransomware virus aimed at mobile phone tended to disguise as auxiliary tools for mobile game, making users install it without any suspicion. Under that circumstance, to some extent, the new mobile ransomware virus distracted users' attention from the initial outbreaking virus. Finally, from duration 5 to duration 15, the topic focus returned to the initial virus and started to concentrate more on the affected regions.

By analyzing the two example topics, we find that the topics discovered by dynamic topic model are explainable enough. Particularly, notable difference can be reflected on the topic evolution process when some certain new events take place.

5. CONCLUSION AND DISCUSSION

In this paper, we study the topic evolution based on the text data from social networking service. To be more specific, we choose keyword “ransomware" as our surveyed object, crawl related tweets from sina Weibo and analyze how the discussion unfolds. After word segmentation and stop words removement, we generate a dictionary and get each tweet vectorized, which is the prerequisite step of topic discovery. Furthermore, by applying dynamic topic model from reference [20], we attempt to analyze how the relative topics evolves as time goes on. At last, we analyze two examples of the resulting topics from the dynamic topic model. It turns out that the experimental outcomes are explainableand, and each transition of the topics can be mapped to a turning of the event in real world. In general, the survey we have done can help users to understand the structure and evolution pattern of event-related topics more thoroughly.

Our future work may lie in several perspectives. First, we will attempt to build a more general model to perceive multiple topics based on probabilistic graphical model. With the event-related topics perceived, we will then try to do the inference online interactively, expecting to realize the early warning for the transforming events before their developments are out of control.

REFERENCES

1. A. Richter, M. Koch, Functions of social networking services, in: Proceedings of the International Conference on the Design of Cooperative Systems 2008, 2008, pp. 87–98.
2. C. Zhang, M. Dong, K. Ota, A social-network-optimized taxi-sharing service, IT Professional 18 (2016) 34–40.
3. G. Dan, We the Media: Grassroots Journalism by the People, for the People, O'Reilly Media, Inc., 2006, pp. 519–548.
4. H. Liu, D. Lee, Quantifying political legitimacy from twitter, in: Social Computing, Behavioral-Cultural Modeling and Prediction, volume 8393, 2014, pp. 111–118.
5. E. G. Gilad Lotan, M. Ananny, The revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions, International Journal of Communication 5 (2011) 1375–1405.
6. W. W. XiaoWei, Daniel Dajun Zeng, Y. Dai, Building the concept semantic space for large text database, Computer Systems Science and Engineering 30 (2015).
7. K. O. Kaimin Wei, Mianxiong Dong, K. Xu, Camf: Context-aware message forwarding in mobile social networks 26 (2015) 2178–2187.
8. M. K. Bergman, White paper: The deep web: Surfacing hidden value, Journal of Electronic Publishing 7 (2001).
9. S.-W. K. Seungdo Jeong, B.-U. Choi, Effective indexing and searching with dimensionality reduction on high-dimensional space, Computer Systems Science and Engineering 31 (2016).
10. P. Barrio, L. Gravano, Sampling strategies for information extraction over the deep web, Information Processing and Management 53 (2017) 309–331.
11. A. Y. N. David M. Blei, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.
12. J. J. Wayne Xin Zhao, J. Weng, Comparing twitter and traditional media using topic models, in: European Conference on Information Retrieval, volume 1, 2011, pp. 338–349.
13. E. A. Yu Wang, M. Benzi, TM-LDA: Effcient online modeling of latent topic transitions in social media, in: Knowledge Discovery and Data Mining, volume 1, 2012, pp. 123–131.
14. R. Fernndez-Beltran, F. Pla, Latent Topic Encoding for Content-Based Retrieval, Springer International Publishing, Cham, 2015, pp. 362–369. doi:10.1007/978-3-319-19390-8_41.
15. H. Liu, D. Lee, Quantifying Political Legitimacy from Twitter, Springer International Publishing, Cham, 2014, pp. 111–118. doi:10.1007/ 978-3-319-05579-4_14.
16. D. Acemoglu, A. E. Ozdaglar, Opinion dynamics and learning in social networks, Dynamic Games and Applications 1 (2011) 3–49.
17. N. E. F. Peng Jia, Anahita Mirtabatabaei, F. Bullo, Opinion dynamics and the evolution of social power in influence networks, Siam Review 57 (2015) 367–397.
18. F. Chen, W. Qu, L. Nie, J. Wu, Y. Li, Discovering probabilistic frequent closed itemsets in uncertain database with tuple uncertainty, Computer Systems Science and Engineering 31 (2016).
19. D. X. Huaping Zhang, Hongkui Yu, Q. Liu, Hhmm-based chinese lexical analyzer ictclas, Sighan Workshop on Chinese Language Processing 17 (2003) 63–70.
20. D. M. Blei, J. D. Lafferty, Dynamic topic models, in: International Conference on Machine Learning, 2006, pp. 113–120.