Advances in Intelligent Systems Research, volume 161
International Conference on Transportation & Logistics, Information & Communication, Smart City (TLICSC 2018)

Hot Events Detection for Chinese Microblogs Based on the TH-LDA Model

Jiahui Chen[1, a], Qingxia Shang[2, b, *] and Hailing Xiong[3, c]

[1]College of Computer and Information Science, Southwest University, Chongqing 400715, China.

[2]College of Computer, Chongqing University, Chongqing 400044, China.

[3]Business College, Southwest University, Chongqing 402460, China.

[a] chenjh9307@163.com, [b, *] xionghl@swu.edu.cn, [c]shangqingxgm@gmail.com

Abstract. Nowadays, many unexpected topics in the society initiate on the microblog platform and spread rapidly, and some of them finally become hot events. The technology for detecting these hot events on the microblog platform, has exerted a very positive influence on the discovery of the latest social hotspots and the timely perception of the internet public opinion. This paper analyzed the studies for detecting microblog hot events, and disclosed the existing methods may omit the untagged microblogs and thus lead to the failure to detect the subevents. To address this issue, we combined Hashtag and Time with the topic model of LDA, and proposed the TH-LDA model to improve the effectiveness of hot events detection for Chinese microblogs. Experiments on microblogs datasets demonstrate that the proposed TH-LDA model can effectively obtain untagged microblogs, and then realize the subevents detection of hot events with a high accuracy.

Keywords: Hot Events Detection; LDA Model; Chinese Microblogs.

1. Introduction

On the microblog platform, information has grown and spread rapidly. Many unexpected topics in the society usually initiated on microblogs [1]. Then users can discuss these topics and share their ideas everywhere and every moment. Consequently, some of such topics become society hot events, which can arouse a wide public response and influence traditional media, to eventually result in great social significance. The detective technology of hot events on the microblog platform has exerted a very positive influence on the discovery of the latest social hotspots. In general, hot events discovery approaches have two categories: unexpected events detection and events discovery [2]. The mechanism of disclosing unexpected events is to build a model on unexpectancy, and to detect the burst points in the information flow by analyzing time series data. The method of events discovery is to detect those events in the real world that develop and evolve in both time and space [8,9].

Twitter is a pioneer of large microblogs, so that many researches on microblog focus on Twitter. The most frequently used method to dig hot events of microblogs is to obtain hot words first, and then identify hot events by these hot words. Kleinberg et al [3] proposed an automaton model to detect text flow by transferring the states of simulative characteristic words and using different states of simulative characteristic words to symbolize their word frequency. As a result, it can reveal the occurrence and disappearance of unexpected events according to different states and word frequency. Takeshi Sakaki et al [4] mainly used the keyword and a Bayesian decision method to detect sudden hotspot events in a specific field. They realized a Twitter-based seismic monitoring system, which achieves a recall rate of 0.8 or more in the application process, and the detection speed is obviously fast than the Seismological Bureau. Tumasjan et al [5] used a text analysis software of LIWC to study the tweets about the German federal elections for exploring the impact of the microblog text on the elections. Their study concludes that the content of the tweets can reasonably reflect the offline political pattern. Becker et al [6] constructed a rule-based classifier that selects a better query strategy from the new query set, to retrieve more new tweets related to the event, and finally realizes the information expansion of the target event.

Textual content of microblogs reflects the enthusiasm and state of the user. At the same time, the release time of the text content is also related to the real behavior characteristics of user [7]. Events tend to evolve over the time, Wang X et al [8] proposed a new LDA topic model by considering the

factor of release time. By using this model, the distribution of dynamic words and the evolution trend of the topic could be obtained. Furthermore, Wang Y et al [7] proposed a TM-LDA model to mine the text stream, simulate the topic transition naturally formed in the data, and obtain the transition parameters of TM-LDA by reducing the prediction error of the topic distribution in the subsequent tweets. After training, the TM-LDA model accurately predicts the distribution of topics in future tweets. Comparing with the work presented in [8], the TM-LDA model prefer to get the relationship among topics. Most of these studies are based on the method of extracting feature words to detect the hot events, and then to explore the evolution of hot events over time. However, Cui et al [9] started from microblogs' "tag" and built an appropriate metric model to reveal the hot events hidden behind it.

Microblogs have many characteristics, such as multi-attribute, colloquial-content, strong noisy. Though the tag attribute has been used to detect hot events in the existing approaches, most of them only consider tagged microblogs, ignoring the untagged microblogs. However, there is a huge number of untagged microblogs on the internet. The microblog collection of the same hot event becomes incomprehensive in the case of ignoring the untagged information. As a result, it will result in the lack of microblog information, and then influence the detection of subevents in hot events. In summary, it will have a great impact on the evolution of tracking hot events.

Motivated by this fact, we proposed the TH-LDA (Time Hashtag Latent Dirichlet Allocation) model, which combines the hashtag and the time factor with the model to better detect the hot events. We then made use of the TH-LDA model to retrieve untagged microblogs, which belong to a same hot event with tag texts, for detecting the subevents of the hot event, so that it is possible to track the evolution of hot events.

2. Proposal for the Tag Classification Algorithm

Tags in microblogs reveal the topics of microblogs. By modeling and analyzing tags, we can dig out hot events hidden behind them. In this section, we proposed the tag classification algorithm to get hot events.

Firstly, we counted the number of different tags which was denoted by n. Each different tag was denoted by h_i and vector H_t was used to denote all k different tags. At the same time, the same tags and all non-null tags were classified to the same h_i. Then, the vector H_t was denoted by:

$$H_t = (h_1, h_2, \cdots, h_k) \quad 1 \le k \le n \tag{1}$$

We then used $(term, w)$ to denote the tag through word segmentation of h_i by using ICTCLAS [10,11]. So, the tag h_i is denoted by:

$$h_{\mathrm{i}} = \{(term_1, w_1), (term_2, w_2), \cdots, (term_{\mathrm{r}}, w_{\mathrm{r}})\} \tag{2}$$

Where $term_{\mathrm{r}}\,(\mathrm{r}=1,2,...,\mathrm{m})$ is term, $w_{\mathrm{r}} = (\mathrm{r} = 1,2,...,m)$ is the weight of $term_{\mathrm{r}}$. Due to the short content of microblog tags, the content of tags can reflect the theme of microblogs directly. Therefore, word frequency could be used to calculate the weight of term, and the weight w_{r} of $term_{\mathrm{r}}$ in tag h_{i} is:

$$w_{\mathrm{r}} = TF_{term_m, h_{\mathrm{i}}} \tag{3}$$

Where TF_{term_m, h_k} is the frequency of $term_m$ occurred in h_k.

Because of the diversity of Chinese expressions, different tags often express the same topic, such as "# Tianjin Binhai Explosion#", "# Tianjin Tanggu Big Bang" and so on. As a result, hashtags with the same meaning must be merged, and the processes are as followed:

Step 1: Count the influence degree of different tags p_{i} in tagged microblogs.

(1) For each tag $h_{\mathrm{i}}\,(\mathrm{i}=1,2,..,\mathrm{n})$, count the number of times that appear in the microblog dataset $n_{h_{\mathrm{i}}}$.

(2) Set each tag h_i (i=1, 2,.., n) 's influence is p_i, $p_i = n_{h_i} \Big/ \sum_{i=1}^{n} n_{h_i}$, and sort it by descending. If the hot event is hotter, the number of microblogs will be larger. So, the impact degree of tag p_i should also be large.

Step 2: Get the threshold of the tag's influence θ.

(1) We solved the mean $c_m = \sum_{i=1}^{n} n_{h_i} \Big/ n$ of all tags n_{h_i} (i=1, 2,..., n) in $[\min\{n_{h_1}, n_{h_2}, ..., n_{h_n}\}, \max\{n_{h_1}, n_{h_2}, ..., n_{h_n}\}]$.

(2)Based on (1), the number of different tags m, their mean $c_{lm} = \sum_{i=1}^{m} n_{h_i} \Big/ m$ in $[\min\{n_{h_1}, n_{h_2}, ..., n_{h_n}\}, c_m)$, the number of different tags $(n-m)$ and their mean $c_{hm} = \sum_{j=m+1}^{n} n_{h_j} \Big/ (n-m)$ in $[c_m, \max\{n_{h_1}, n_{h_2}, ..., n_{h_n}\}]$ should be solved separately.

(3) On the basis of different scale data sets we got that if $(c_{hm} - c_m)/(c_m - c_{lm}) \le 8$, then $\theta = \min\{n_{h_1}, n_{h_2} ..., n_{h_n}\} / n$. Otherwise we should continue to solve the mean of the total tags in $[c_m, \max\{n_{h_1}, n_{h_2} ..., n_{h_n}\}]$ until c_θ satisfies the condition $(c_{h\theta} - c_\theta)/(c_\theta - c_{l\theta}) \le 8$. Where $c_{h\theta}$ is the mean in $[c_\theta, \max\{n_{h_1}, n_{h_2} ..., n_{h_n}\}]$ and $c_{l\theta}$ is the mean in $[c_m, c_\theta)$.

Step 3: According to the threshold θ in Step 2, we obtained all the tags satisfied $p_i > \theta$.

Step 4: We segmented the chosen tags and counted word frequencies, then utilized cosine similarity to classify them and got the hot events set $E_{set} = \{e_1, e_2, ..., e_q\}$, where $E_{set} = \{e_1, e_2, ..., e_q\}$ denote the different events. At the same time, we can obtain tag sets $EH_{set_j} (j = 1,2,...,q)$ of every hot event.

Based on the tag set $EH_{set\,j}$, we obtained all hashtags and all tagged microblogs of every hot event e_i. Then the tagged microblogs can be stored in the collocation TE_i, and we used terms to represent the category of hot events, $i(i \le n)$ denotes the hot event. In this paper, the high-frequency vocabulary in the tagged microblogs of the event is selected as topic keyword $term_j$ to mark the event. The theme term vector of the hot event e_i is expressed as:

$$e_i = (term_1, term_2, \cdots, term_j) \tag{4}$$

The tag classification algorithm proposed in this section can be used to detect hot events of microblogs. However, many users are most likely to express their opinions on a hot event without tags. Table 1 shows the proportion of total 134,0461 microblogs with and without tags in microblogs. It can be seen from Table 1 that the number of untagged microblogs is 76.97%, which is much larger than the proportion of tagged microblogs, which is 23.03%. Therefore, obtaining a more complete microblog collection containing tags is necessary for hot event detecting.

Table 1. The proportion of total microblogs with and without tags

Tag	Text Volume	Proportion
Tagged Microblogs	308706	23.03%
Untagged Microblogs	1031755	76.97%

Tag classification algorithm proposed in this section can only detect hot events by hashtags straightforward. Yet one hot event usually evolves and develops many subevents, which play an important role in fueling the evolution of hot events. To better detect subevents and track the evolution of events, it is also necessary to analyze subevents.

3. Hot Event Detection based on TH-LDA Model

3.1 Proposal for the TH-LDA Model

Based on section 2, we combined the event time factor (Time) to obtain the untagged microblog collection in the same hot event, and merged it with the tagged microblog collection obtained to get a complete collection of microblogs. Also, we proposed the TH-LDA model to detect subevents and track the evolution of events.

Time is the carrier of generation, development and extinction of events. Therefore, the amount of microblogs often experience a process from rise to fall over time. To reduce the redundancy of microblogs, we used the time of the tagged microblogs to limit the hot event time after introducing the tags.

Suppose that the number of tagged microblogs that belongs to a hot event is n.

Definition 1(Event start time) The event start time is the earliest time of posting the tagged microblog related to this event, recorded as $t_{o_i}(i \le n)$.

Definition 2(Event end time) The event end time is the latest time of posting the tagged microblog related to this event, recorded as $t_{e_i}(i \le n)$.

Definition 3(Event duration interval) The event duration interval is from the start time of the event to the end of the event, which is recorded as T_i or $T_i = [t_{o_i}, t_{e_i}](i \le n)$.

The tags related to the hot event is obtained according to the tag classification algorithm, and all the microblogs in the event duration interval are obtained. On this basis, steps of constructing the TH-LDA model as follows:

Step1: Get all the collections of untagged microblogs in the interval T_i, and use Vector Space Model (VSM) to represent them. The process as follows:

The event duration interval is used as the time slot limit, and collect all untagged microblogs in the event duration interval to make up microblog collection TW_i. Each microblog in TW_i is represented as a text vector tw_j. Use tm_k and wm_k to represent tw_j's each the terms and their corresponding weights, the number of which depends on the number of words in each microblog, then the text vector tw_j is expressed as:

$$tw_j = \{(tm_1, wm_1), (tm_2, wm_2), \cdots, (tm_k, wm_k)\} \tag{5}$$

wm_k is the weight of tm_k in microblog collection TW_i, tm_k is key word of microblog tw_j. The term frequency is used and the maximum normalization of it:

$$wm_k = TF_{tw_j, tm_k} \Big/ \max_{tm_k \in tw_j} \{TF_{tw_j, tm_k}\} \tag{6}$$

Where TF_{tw_j, tm_k} is the term tm_k's word frequency in microblog tw_j.

Step2: Get all the microblog collections TS_i that do not contain tags in the same hot event:

Using the topic vector of hot event obtained in in Section 2, perform keyword matching query on all the untagged microblog vectors tw_j. Then obtained all the untagged microblogs vectors ts_j that belong to a certain type of event e_i, and stored them in the microblog collection TS_i.

Step3: The tagged microblog collection TE_i and the untagged microblog collection TS_i in each type of hot event are merged into the microblog collection TS_i.

Step4: For each hot event in microblog collection TES_i, use T_p as time interval to segment into $m = \lceil T_i / T_p \rceil$. Using the LDA model, the m parts chronological files are online trained during the total time interval T_i.

4. Hot Event Detection

After obtaining a complete microblog collection, we detected subevents of hot events to track the evolution of events. The algorithm steps of hot event detection based on TH-LDA model are as follows:

Step1: Use the TH-LDA model to train hot events to obtain the file words of topic-word distributions.

Step2: The words file is analyzed to obtain the topics in each time slot T_p , then analyze the topics and terms belong to the topics to obtain the subevents. The text vector representation denotes the subevent topics, and the cosine similarity is used to obtain the microblog collection of the same subevent. The pseudo code of hot event detection based on TH-LDA model are as follows:

```
getTHLDA () {
  To = getEventMinTime (); // Earliest start time of the event
  Te = getEventMaxTime (); // Event end time
      Ti= Te. getTime () - To. getTime();    // Time interval of getting events
  m = Math. ceil (Ti/Tp); // Splitting the event into m portions
  gettextset();    /Gets all the microblog text sets in the event interval Ti.
  getTplit();    // Word segmentation of microblog text
  for 1 to m
  {
       Gettxt (); // Get each chronological data and save it in a .txt file
       LDA (); //Train LDA model
            getSimilarDocuments (); // Aggregate based on similar documents at the
topic level
  }
}
```

5. Experiment and Analysis

To simulate the real microblog environment, we used redundant microblogs as experimental data. The TH-LDA model detects unknowns from unknowns, so it cannot be evaluated by precision, recall and F-measure. Therefore, we only tested the feasibility of the model and did not compare it with the topic model that only detects hot events.

5.1 Experimental Data Set

The experimental data set is derived from randomly crawling the content of Sina microblogs' public webpage, consisting of 2231, 8948 microblogs from August 12, 2015 to August 20, 2015. As can be seen from Figure 1, the number of microblogs is between 1.5 million and 3.5 million, and the amount of average daily microblogs is 247, 9904. The total amount of data tagged per day is between 20,000 and 500,000, and the average daily tagged microblogs is 34,457.

This paper collected 310, 1129 tagged microblogs as the data set of the tag classification algorithm. Firstly, 180, 5243 microblogs of the Tianjin Explosion were collected by the tag classification algorithm for training the TH-LDA model. Secondly, 85,123 microblogs of the 70th anniversary of the victory of the Chinese People’s War of Resistance against Japanese Aggression were obtained as the test data set of the TH-LDA model.

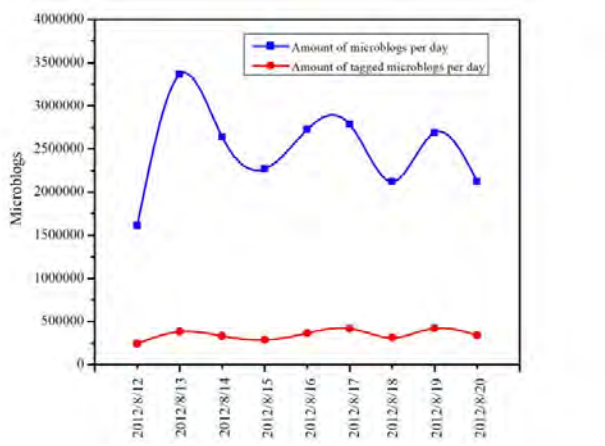


Figure 1. Distribution of microblog data in experimental data sets

5.2 Parameter Settings

Set $t_{o_i} = 2015\text{-}08\text{-}12\ 13{:}46{:}39$, $t_{e_i} = 2015\text{-}08\text{-}20\ 15{:}59{:}08$ and $T_i = [2015\text{-}08\text{-}1213{:}46{:}39, 2015\text{-}08\text{-}2015{:}59{:}08]$. Set $T_p = 24h$ and divide event database into $m = \lceil T_i / T_p \rceil = 9$, use the 9 pieces of data to train the model.

Set the number of topics $K = 30$, parameters $\alpha = 0.5$, $\beta = 0.01$. The model was trained by the Gibbs sampling algorithm and reached the convergence after 2000 iterations.

5.3 Analysis of the Experiment and Results

5.3.1 Tag Classification

There are 441,446 different tags in tagged microblogs. Tags reflects the topic of microblogs, so we defaulted that the tagged 3,101,129 microblogs contained 441,446 topics. The number of different tags that satisfy the influence threshold P is 2695 by using tag classification algorithm.

By using the tag classification algorithm to classify 2695 different tags, 447 categories were obtained. From this category set, we got 26 categories included hot events after eliminating redundant categories such as entertainment and marketing promotion. For some hot events (such as "# tribute to firefighters" in "Tianjin Explosion" and "#Tangyu Explosion Truth #"), although they belong to different categories, they describe different the same hot event. After merging them, a total of 13 hot events were obtained, as shown in Table 2.

From detected hot events, the microblog collection of the Tianjin Explosion was chosen as the training data set of the TH-LDA model. Taking the day as time intervals, the training data set is divided into 9 parts, which is in the time range of $T_i = [2015\text{-}08\text{-}12\ 13{:}46{:}39, 2015\text{-}08\text{-}20\ 15{:}59{:}08]$, from August 12 to 20. Figure 2 shows the change in total daily data over the time range of the Tianjin Explosion.

As can be seen from Fig.2, the total amount of microblogs about the Tianjin Explosion increased sharply on August 13. This is consistent with the sudden and rapid spread of hot events. According to official reports, Tianjin Explosion occurred at 23:30 on August 12, 2015. The start time of the Tianjin Explosion detected in microblogs was 10 hours earlier than official release. By the end of the official release, the number of microblogs of Tianjin Explosion was 10,9616. This also shows that the hot event is firstly issued on the microblog platform. Therefore, detecting the hot events in microblog and track its evolution and development has certain practical significance.

The chronological 9 data sets were trained by the model. After training, the theme-words (words) files of each data set were obtained and the subject words of the 9 topic-word files were summarized and removed duplication.

Table 2. Hot events obtained using the tag classification algorithm

Hot Event	Number of Different Tags	Total Amount of Tags
E1 Tianjin Explosion	807	173869
E2 70th Anniversary of the Victory of the Chinese People's War of Resistance Against Japanese Aggression	34	6351
E3 Bangkok Explosion	23	1907
E4 Shanyang Landslide in Shanxi	3	690
E5 Anshan Explosion	1	227
E6 the CUC Girl Was Killed	3	461
E7 Fuding Flood	3	1061
E8 Shanghai Elevator Folder Event	1	213
E9 Opening Fire on the Korean-Korean Border	1	424
E10 Indonesian Aircraft Lost	2	377
E11 The Boy was Cut More than 30 Times by His Stepmother	2	362
E12 Siblings Were Feed Sulphuric Acid by Older	2	333
E13 Seven Knives in the Body of a Young Mother were Identified as Suicide	1	248

5.3.2 Hot Event Detection based on the TH-LDA Model

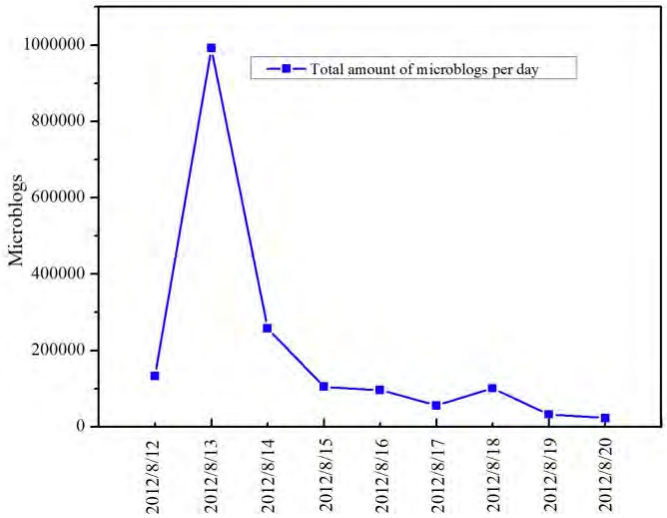
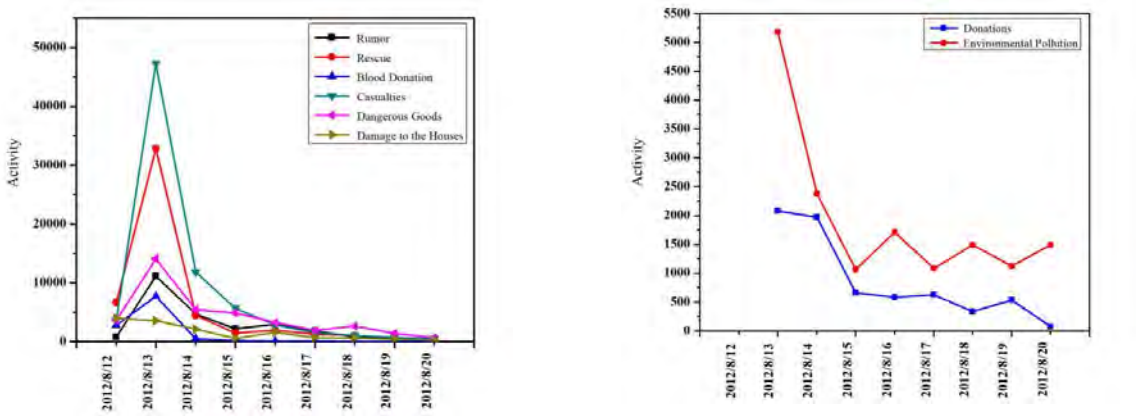


Figure 2. Total amount of microblogs of Tianjin Explosion

Based on the model, the topic-word file in each data set is clustered to similar microblogs and then counted the total amount of similar microblogs in the same topic in each data set. Since each data set is performed daily, we only need to count the total number of daily statistics for the same topic.

The evolution of the event will produce many topics, and the evolution of the topics will most reflect the evolution of the event directly. The number of microblogs on the same topic is used to measure the activeness of the topic. Topics which are throughout the period of the Tianjin Explosion are analyzed. Figure 3(a) shows how activeness of topics such as "rescue", "rumors" and "casualties" changed over time. According to Fig.3(a), the activeness of these topics declined gradually over time, which indicates that people get less concerned about the event and the focus of their attention was changing. For the topic of "rumor", the activeness on August 16 has increased compared with the day before. In the test of the TH-LDA model, it was found that on August 16th, the rumor of "urban management to grab volunteers" appeared, so there will be an increase in activeness.



(a) (b)

Figure 3. How activeness of topics changed over time

During the evolution process of events, several subjective topics will be generated, such as "donations" and "environmental pollution" detected after Tianjin Explosion happened. As can be seen from Fig.3(b), the two subevents were detected on August 13th, and there has been a wave of fluctuations as time went by. According to the TH-LDA model, the reason why the activeness of "donations" fluctuates is because two hot subevents are generated during the period. Table 3 shows the two subevents detected and the total amount of microblogs it occupies in the Tianjin Explosion data set.

Table 3. Subevents detected using the TH-LDA model

Subevent	Total Amount of Microblogs
E1.1 A Guangxi post-1995 Generation Female Lied about the Death of Their Father for Charity Fraud.	8150
E1.2 Jack Ma Forced to Donate	1446

There are many topics in the development of each subevent, and the evolution of these themes can directly reflect the evolution of subevents. Table 4 shows the evolution of the topic "A Guang-xi post-1995 generation female lied about the death of their father for charity fraud." (referred to as "the charity fraud").

Table 4. The evolution of the topic of the charity fraud

Detected Time	Related Topic	Subject Words and Their Probabilities
August 14	Charity fraud of father died	Liar, donation, fraud, alarm, Account number, father, freeze
	Using national calamity and sympathy to make money	Conscience, disaster, national affliction, liar, Compassion, scam, parents
August 15	Guangxi female of charity fraud was under arrest	Netizen, fraud, father, suspected, Guangxi, detention, death
	Adults are responsible for the act and can be sentenced	Responsible, behavior, adulthood, Using Compassion, make use of, sentence

From Table 4, we can draw the evolution of the charity fraud: from discovering "charity fraud" to "using sympathy and deception", then publishing the fraudulent personnel as "post-1995 generation Guangxi girls". Eventually, the defrauder was under arrest and the citizens asked for sentence.

6. Conclusion

This paper proposed a tag classification algorithm to detect the hot events in microblogs. Then combined Hashtag, Time and the LDA model to propose the TH-LDA model. This model is used to retrieve untagged microblogs of the same hot event, and get a more complete microblog collection of the hot event. At last we realized the detection of subevents, and tracked the evolution and development of events. The evaluation experiments have verified the feasibility of the newly proposed model.

Microblogs has the characteristics of informational diversity, which deserves further research in many aspects. In the detection of hot events, we only used tags and time factor of the users. To better utilize the geographical location characteristics of the microblogs for discovering more comprehensive hot events is a direction of future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.41271292). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation. Thanks, are also due to Qingxia Shang from Chongqing University and Hailing Xiong from Southwest University for their field supports.

References

[1]. He Min, Du Pan and Zhang Jin et al, Microblog Bursty Topic Detection Method Based on Momentum Model. Journal of Computer Research and Development, 52(5):1022-1028(2015).

[2]. Cui Anqi, “Study on Public Sentiment Analysis of Events in Microblogs,” Ph.D. thesis, Tsinghua University, 2013.

[3]. Kleinberg J, Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery,7(4):373-397(2003).

[4]. Sakaki T, Okazaki M, Matsuo Y, “Earthquake shakes Twitter users: real-time event detection by social sensors,” in Proceedings of the 19th international conference on World Wide Web, edited by Michael Rappa et al. (World Wide Web, Raleigh, USA,2010), pp.851-860.

[5]. Tumasjan A, Sprenger T O, Sandner P G, et al. “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment,” in International Conference on Weblogs and Social Media, edited by Isabell M, Welpe. (The International AAAI Conference on Web and Social Media, Washington DC, USA, 2010), pp.919-931.

[6]. Becker H, Chen F, Dan I, et al. “Automatic Identification and Presentation of Twitter Content for Planned Events,” in International Conference on Weblogs and Social Media, edited by Ian Soboroff. (The International AAAI Conference on Web and Social Media 2011, Barcelona, Spain, 2011), pp.2011.

[7]. Wang Y, Agichtein E, Benzi M. “TM-LDA: efficient online modeling of latent topic transitions in social media,” in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, edited by Thanawin Rakthanmanon. (Association for Computing Machinery, Beijing, China,2012), pp.123-131.

[8]. Wang X, Mccallum A. “Topics over time: a non-Markov continuous-time model of topical trends,” in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, edited by Thorsten Joachims. (Association for Computing Machinery, 2006), pp.424-433.

[9]. Cui A, Zhang M, Liu Y, et al. “Discover breaking events with popular hashtags in twitter,” in ACM International Conference on Information and Knowledge Management, edited by Thanawin Rakthanmanon. (Association for Computing Machinery, Beijing, China, 2012), pp.1794-1798.

[10]. Zhang H P, Yu H K, Xiong D Y. “HHMM-based Chinese lexical analyser ICTCLAS, Sapporo, Japan: Associ-ation for Computational Linguistics,2003:184-187.

[11]. LiuQun, Zhang Hua-Ping, Yu HongKui et al, “Chinese Lexical Analysis Using Cascaded Hidden MarkovModel”. (Journal of Computer Research and Development, 2004,41(8)), pp.1421-1429.