An Event Extraction Model based on Timeline and User Analysis in Latent Dirichlet Allocation

Bayar Tsolmon
Division of Computer Science and Engineering
Chonbuk National University
Republic of Korea
bayar_277@yahoo.com

Kyung Soon Lee
Division of Computer Science and Engineering, CAIIT
Chonbuk National University
Republic of Korea
selfsolee@chonbuk.ac.kr

ABSTRACT

Social media such as Twitter has come to reflect the reaction of the general public to major events. Since posts are short and noisy, it is hard to extract reliable events based on word frequency. Even though an event term appears in a particularly low frequency, as long as at least one reliable user mentions the term, it should be extracted. This paper proposes an event extraction method which combines user reliability and timeline analysis. The Latent Dirichlet Allocation (LDA) topic model is adapted with the weights of event terms on timeline and reliable users to extract social events. The reliable users are detected on Twitter according to their tweeting behaviors: socially well-known users and active users. Reliable and low-frequency events can be detected based on reliable users In order to see the effectiveness of the proposed method, experiments are conducted on a Korean tweet collection; the proposed model achieved 72% in precision. This shows that the LDA with timeline and reliable users is effective for extracting events on the Twitter test collection.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, information filtering.*

General Terms

Algorithms, Experimentation.

Keywords

Event Extraction; Timeline Analysis; User behaviors; Latent Dirichlet Allocation

1. INTRODUCTION

In recent years, media has become an important source of real-time information. People use social media to communicate, socialize, debate and engage in arguments. With the rapidly increasing amount of data on social networking service (SNS) like Twitter, researches on event extraction are attracting more and more attention. Compared with news and blog data, SNS data are more widely used in real-time event extraction. However, as the amount of data increases, noise also increases, creating a need for a reliable event extraction method.

There are researches on effectively analyzing diversity of tweets and inferring the occurrence and magnitude of an event [2, 3, 5]. Since event extraction methods [6, 7] typically depend on term frequency, it is possible to miss important information when extracting events from Twitter. This is especially true if retweets are not pertinent to an event, such as a rumor, spam or advertisement, all of which are common in the SNS environment. In order to reliably extract reliable low-frequency events as well as high-frequency events, two types of reliable users should be included in the event extraction method: the active user and the socially well-known user. Active users are valuable users because they post important information every time an event occurs. On the other hand, if a socially well-known user mentions a particular event, it indicates that a significant social event occurred. In addition, the events themselves have important features. The day a new event occurs, certain terms may be used more frequently that day compared to other days. Users will also use Twitter to express positive or negative opinions about particular issues or events. Our model uses a sentiment lexicon to consider the context of an event term which contains users' opinions.

A number of studies have been conducted on various forms of social media. Tinati et al. [8] developed a model based on Twitter message exchanges which makes it possible to analyze conversations about specific topics and identify key players in the conversation. Sayyadi et al. [7] developed a new event detection algorithm creating a keyword graph and using community detection methods analogous to those used in social network analysis in order to discover events.

The Latent Dirichlet Allocation (LDA) topic model [1] provides a principled way to discover hidden topics from large document collections. However, standard topic models do not consider temporal information. In the recent research on LDA, the topic model [2] considers both the temporal information of microblog posts and users' personal interests.

This paper proposes an event extraction model, which is a Latent Dirichlet Allocation topic model based on timeline and user behavior analysis. Timeline and user reliability analysis are applied to the standard LDA model [1] to cover multiple events occurring on the same day. A sudden increase in topically similar posts usually indicates an event. Event terms are extracted using the chi-square test and opinion scores to show how term distribution is related to timeline and user sentiment about a specific event. The reliable users are detected on Twitter according to their tweeting behaviors: socially well-known users and active users. The proposed model is novel in that it combines user reliability and timeline analysis to extract event topics. To see the effectiveness of the proposed model, experiments are conducted on a Korean Twitter test collection. The paper is organized as follows: Section 2 presents our event extraction model; Section 3 describes experimental results. Finally, we conclude in Section 4.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright © 2014 ACM 978-1-4503-2257-7/14/07…$15.00.

2. EVENT EXTRACTION MODEL

This section describes the overall structure of the proposed method as shown in Figure 1. The method consists of a topic discovery step and an event filtering step. In the topic discovery step (section 2.1), a topic model is proposed that considers both timeline (section 2.1.1) and user reliability (section 2.1.2). Event filtering is performed with a similarity-based clustering and user reliability-based filtering method (section 2.2).

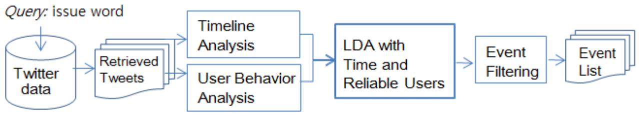

Figure 1. System Architecture

2.1 Event Term Extraction Based on Timeline and Reliable Users in LDA model

This section describes the event term extraction method based on the LDA topic model with time and users. The purpose of using the LDA model is to cover multiple events occurring on the same day. This is in contrast to basic event extraction methods, which are based on term frequency and therefore cannot extract multiple events occurring on the same day. Timeline and user behavior analysis are critical properties in the event extraction method. When a new event occurs on a certain day, there is typically a large occurrence of certain of certain terms on that day compared to other days. In order to give distinctive weight to an event term, term significance can be calculated based on timelines. In addition, when an event occurs, generally public is informed about it by authority users, or through mass media, or through normal users on Twitter. Each time an event occurs there are some users who tend to write profusely about the issue. Such users are concerned about the issue for a long time and may publish valuable information on that issue. Thus, it is important to detect accurate and reliable users in order to extract accurate and reliable events.

In the proposed model, the LDA model is adapted with timeline and user reliability analysis to extract event topic groups (Figure 2). Here, *T* represents time series, *U* indicates user sets, χ indicates the additional weight of each term on time *t*, and π denotes the additional weight of each user. The topic distribution of each user θ is drawn from a prior Dirichlet distribution Dir(α), and each document word w is sampled from topic-word distribution ϕ specified by a drawn from the topic-user distribution θ.

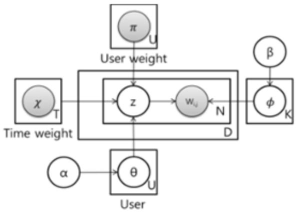

Figure 2. Graphical representation of TimeReliableUser LDA

Collapsed Gibbs sampling was used to perform model inference. Due to space limitations, only the derived Gibbs sampling formulas are shown, as follows:

$$p(z_i = j|w, z_{-i}) = \frac{ChiOpScore(m)C_{mj}^{WK}+\beta}{\sum_{m'} C_{m'j}^{WK}+V\beta} \cdot \frac{AuthScore(n)C_{nj}^{UK}+\alpha}{\sum_{j'} C_{nj'}^{UK}+K\alpha} \quad (1)$$

where $z_i = j$ represents the assignments of the *i*th word in a document to topic *j*. z_{-i} represents all topic not including the *i*th word. *ChiOpScore* is the additional weight of *m*th word in the lexicon on time *t*. *AuthScore* denotes the additional weight of *n*th user. The range of *ChiOpScore* and *AuthScore* are (0,1]. Furthermore, C_{mj}^{WK} is the number of times word *m* is assigned to topic *j*, not including the current instance, and C_{nj}^{UK} is the number of times user *n* is assigned to topic *j*, not including the current instance. The proposed model is similar to the model proposed by Diao *et al*. [2]. The primary difference between the two is that, they assume a pair of posts published around the same time is more likely to be about the same topic than a random pair of posts. Here, term significance for a given day is measured by the chi-square statistic, which measures the lack of independence between a term and the date. Another difference is that they compare a post with its publisher's general topical interests observed over time. If a post does not match the user's long-term interests, it is more likely related to a global event. In this paper, reliable users are concerned to extract an event. Then, if the reliable users post about the issue on a particular day, it is more likely related to an event. In addition, the model by Diao *et al*. extracts an event by two Poisson distribution. The proposed method extracts social events based on topic term clustering and reliable users based filtering.

2.1.1 Timeline Analysis

When a new event occurs on a certain day, there may be a high occurrence of certain terms on that day compared to other days. The term significance is measured by the Chi-square statistic, which measures the lack of independence between a term and the date. Users write tweets to express their opinions on particular issues or events with positive or negative sentiment words. A sentiment lexicon is used to detect the context of an event term which contains the user's opinion.

Event terms are extracted by combining the Chi-square value (*ChiSq*) and opinion score (*OpScore*) to show how term distribution is related to the timeline and user sentiments for a specific event, as follows (see more details in [9]):

$$ChiOpScore(w, t_0) = \lambda \cdot ChiSq(w, t_0) + (1 - \lambda) \cdot OpScore(w, t_0) \quad (2)$$

where *w* is a bigram word, t_0 is a particular date, and λ is set to 0.7 empirically. *OpScore* is measured by combining the frequency of a term and the frequency of opinion words. The *ChiOpScore* value (additional time weight χ) of each term is applied to the TimeReliableUser LDA model to account for the impact of event terms.

2.1.2 User Behavior Analysis

When an event occurs, generally the public is informed about it by authority users, or through mass media, or through normal users on Twitter. In addition, each time an event occurs there are some users who tend to write profusely about the issue. Such users are concerned about the issue for a long time and may publish valuable information on that issue. For example, weather information from the official Twitter account of the National Weather Service is accurate and reliable. Similarly, it can be quick and reliable to get the latest information about iPhones from the official Twitter account of Apple. Thus, it is important to detect accurate and reliable users in order to extract accurate and reliable events. A critical part of the event extraction model is the classification of Twitter users. In our model, users are categorized as either socially well-known users or active users. If a socially well-known user mentions a particular event, it indicates that an important social event occurred. Highly active users, on the other hand, post important information whenever a new event occurs. Both of these groups are considered reliable users.

2.1.2.1 Detecting socially well-known users

Generally, socially well-known users on Twitter tend to have a large number of tweets and retweets. A HITS algorithm [4] was adapted to extract socially well-known users by applying mentions, RTs (modified retweet), and retweets as an edge weight between user nodes.

$$AuthScore^{(T+1)}(p) = \sum_{q \to p} w_{qp} \times HubScore^{T}(q) \qquad (3)$$

$$HubScore^{(T+1)}(p) = \sum_{p \to q} w_{pq} \times AuthScore^{T}(q) \qquad (4)$$

The edge weight w_{qp} is as follows:

$$w_{qp} = \sum_{q \to p} FreqRT(q,p) + \sum_{q \to p} Mention(q,p) \qquad (5)$$

The 5 users with the highest *AuthScore* values are considered socially well-known users.

2.1.2.2 Detecting active users

These users generally post profusely about a given topic and related events, and thus are typically more active than other users. The following formula calculates the average weekly activity score.

$$Activity\ Score(u) = \frac{1}{W}\sum_{i=1}^{W} TweetFreq(u,d_i) \times RTFreq(u,d_i) \qquad (6)$$

where *W* show the number of weeks; *TweetFreq* shows the sum of tweets *d* that a user *u* wrote in the *i*th week; *RTFreq* represents the number of retweets *d* of the tweets written by a user *u* in the *i*th week.

The distribution of reliable user tweets about earthquakes and the number of news articles about earthquakes on a given date are shown in Figure 3. The number of tweets about an earthquake, the tweet distribution of the top 5 reliable users, and the number of news articles about the earthquake all have very similar patterns. Peaks on the graph labeled by date represent the occurrence of earthquakes. An increase in daily tweet frequency indicates that all reliable users wrote about an issue on the same day.

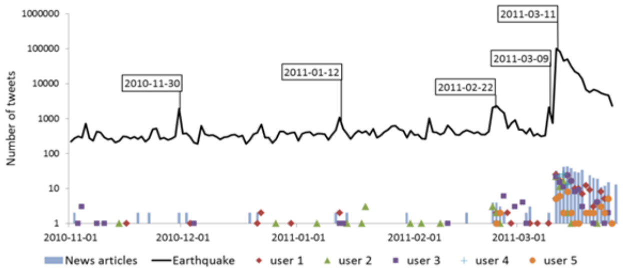

Figure 3. The number of tweets containing a word "earthquake," the distribution of reliable users and the number of earthquake related news articles.

The top 5 reliable and highly active users for the term "earthquake" are shown in Figure 3. User 1(Korean Red Cross) wrote tweets not only about the earthquake itself, but also about ways to donate or help. In contrast, user 5 (Ministry of Foreign Affairs) posted tweets that include guidelines for South Korean citizens who live in abroad, while users 2 (KBS news) and 3 (journalist) provided earthquake-related news.

The five users with the highest activity score are selected as active and reliable users. Both socially well-known users and active and reliable users are classified as reliable users. These reliable users are selected as input data π for the our LDA model.

2.2 Event Filtering

The results of the TimeReliableUser LDA model can include similar topic groups and noisy, high-frequency event terms which are not related to the event. Thus, the extracted events are filtered by user data from reliable and highly active users. Ideally, this process will give a lower rank to highly frequent but unrelated event terms, and a higher rank to frequent but important event terms. The system creates 20 topic groups for each day and each topic group has 20 words and 20 users. Duplicate topic groups were removed by applying the cosine similarity measure, which measures the similarity between two vectors by finding the cosine of the angle between them. The process is as follows:

Step 1: Compute similarity score between each topic-word group
Step 2: Compute similarity score between each topic-user group
Step 3: If the average similarity scores of each topic-word pair and topic-user pair are greater than the threshold, then group them together (single link clustering).
Step 4: Select a topic-word group as an event group if a topic-user group includes reliable users.

The threshold of cosine similarity for clustering topic-word groups was set at 0.5 based on results from the qualitative analysis and analysis of the number of topic pairs. The event topic group filtering process is shown in Figure 4.

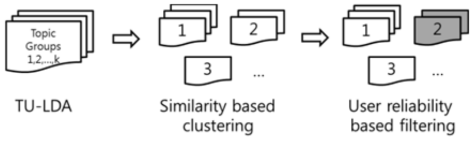

Figure 4. Topic clustering and filtering process

In Figure 4, the first step provides several topics and each topic group has an associated probability score. The number of topic group sets is obtained after measuring cosine similarity between the topic groups. Finally, the topic groups are filtered by reliable users or author reliability score. Even if an event term has a particularly low frequency of occurrence, as long as at least one reliable user mentioned the term, it can be extracted.

3. EXPERIMENTS

3.1 Experimental Setup

The effectiveness of the proposed method of tweet collection was evaluated as follows: Four topics were selected, and tweets related to these topics or events were collected from November 1, 2010 to March 26, 2011 by Twitter API (all tweets are written in Korean). Table 1 shows the number of tweets related to each issue and 41 social events for the four issues. Each record in this tweet data set contains the actual tweet body and the time when the tweet was published.

Table 1. Korean Twitter data set

Issues	*# of users*	*# of tweets*	*# of events*
Park Ji-Sung	29,568	131,533	12
Kim Yu-Na	10,563	26.844	8
Earthquake	110,345	467,955	10
Cheonanham	19,473	84,195	11
Total	169,949	683,710	41

First, the model filtered out stop words and extracted nouns and verbs using the Korean POS(part-of-speech) tagger. After preprocessing, the proposed method created a word pair of features in each tweet by using a bigram with 3 window size.

3.2 Experimental Results

Comparison methods for event extraction are as follows:

- **Standard LDA:** Original topic model[1] [1]
- **BurstTimeUser LDA:** LDA with time and users to find bursty topics [2]
- **TimeReliableUser LDA:** The proposed model based on reliable users and timeline analysis

Two human assessors judged the answers for the result topics by each method on the particular dates which an event occurred. The ten dates in average are selected according to the number of news articles on the issue. After preliminary experiments, the number of topics was set at K to 20, α to 0.01 and β to 0.01, empirically. The TimeReliableUser LDA was run for 500 iterations of Gibbs sampling, with time T spanning from November 1, 2010 to March 26, 2011 for a total of 147 days.

Table 2. The results of the comparative experiments

Method	Precision
Standard LDA	0.55
BurstTimeUser LDA	0.63
TimeReliableUser LDA	0.72

Table 2 shows experimental results on 41 events, with the number of answers detected out of the total number of answers for each event. Evaluation process is as follows. Each model gives 20 topic groups on particular day. Here, the proposed model gives clustered topic groups with reliable user groups. When the number of tweets is less than other day, the proposed model gives two or three topic groups. And when the numbers of tweets are high then the model gives some topic groups which have similar user groups. The baseline LDA model achieved 55%, BurstTimeUser LDA achieved 63%, and the proposed model, TimeReliableUser LDA, achieved 72% in macro average precision. Our model outperforms other models in precision and recall. This shows that timeline and reliable users played an effective role on event extraction.

We show some sample results from our experiments and discuss some case studies that illustrate the advantages of our model. Examples of the extracted event groups are shown in Table 3. The topic labels are manually assigned. Each topic group consists of top words and top users. Each topic is shown with five words and users that have the highest probability conditioned on that topic. As shown in Table 3, the first and second topic groups are event groups because those groups include reliable users. Third topic groups are also event groups, but the system could not retrieve them. The reason is that groups are not including reliable users. However similarity-based clustering gave us meaningful topic groups and it reduced the topic groups depending on the event type.

Table 3. Example topics on "Japan earthquake"

Topics	*Top words*	*Top users*
Japan earthquake (clusters 0, 2, 5, 6, 12)	Japan earthquake Earthquake scale Magnitude 8.9 Tsunami warning Japan tsunami	**National Weather Service** **South Korean embassy in Japan** Normal user Normal user Normal user
Free phone service (clusters 13, 15, 18)	Family connections The earthquake area Contact local residence area Consular Call Center Free operators	Normal user **Blue House Korea** **Ministry of Foreign Affairs** Normal user **Korean Red Cross**
Radiation leakage (clusters 14, 17, 19)	Earthquake in Japan Japan earthquake Radiation leakage Nuclear Radiation Earthquake Information	Normal user Normal user Normal user Normal user Normal user

4. CONCLUSION

This paper proposed an event extraction method based on a timeline and user behavior analysis in LDA on Twitter. On the TimeReliableUser LDA model, event candidate terms and reliable users were reflected by term significance and user activity measurements. Similar topics are clustered by measuring term relevance and by considering reliable users. The proposed method achieved 92% average precision in the top 10 results. The study findings show that using the TimeReliableUser LDA model with reliable users can increase the effectiveness of event extraction.

Future research should focus on methods for better event expression and less dependency on the number of tweets.

5. ACKNOWLEDGMENTS

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A1A2044811).

6. REFERENCES

[1] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet Allocation. The Journal of Machine Learning research 3, pp. 993-1022

[2] Diao, Q., Jiang, J., Zhu, F., and Lim, E.P. 2012. Finding bursty topics from microblogs. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 536-544.

[3] Kanhabua, N., and Nejdl, W. 2013. Understanding the diversity of tweets in the time of outbreaks. In Proceedings of the 22nd international conference companion on World Wide Web, pp. 1335-1342.

[4] Kleinberg, J. M. 1999. Authoritative Sources in a Hyper-linked Environment. Journal of the ACM, 46(5), pp. 604-632.

[5] Lampos, V., and Cristianini, N. 2012. Nowcasting Events from the Social Web with Statistical Learning. ACM Transactions on Intelligent Systems and Technology 3(4:72).

[6] Popescu, A. M., and Pennacchiotti, M. 2010. Detecting Controversial Events from Twitter. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp.1873-1876.

[7] Sayyadi, H., Hurst, M., and Maykov, A. 2009. Event Detection and Tracking in Social Streams. In Proceedings of 3rd AAAI International Conference on Weblogs and Social Media, pp. 311-314.

[8] Tinati, R., Carr, L., Hall, W., and Bentwood, J. 2012. Identifying Communicator Roles in Twitter. In Proceedings of the 21st International conference companion on World Wide Web, pp. 1161-1168.

[9] Tsolmon, B., Kwon, A., and Lee, K.-S. 2012. Extracting Social Events Based on Timeline and Sentiment Analysis in Twitter Corpus. Lecture Notes in Computer Science, vol. 7337, pp. 265-270.

[1] Implementation of GibbsLDA http://gibbslda.sourceforge.net/