SCIENCE CHINA
Information Sciences

• LETTER •

December 2018, Vol. 61 129101:1–129101:3
https://doi.org/10.1007/s11432-016-9333-4

An event summarizing algorithm based on the timeline relevance model in Sina Weibo

Kai LEI*, Lizhu ZHANG, Ying LIU, Ying SHEN, Chenwei LIU, Qian YU & Weitao WENG

Shenzhen Key Lab for Cloud Computing Technology & Applications (SPCCTA), School of Electronics and Computer Engineering, Peking University, Shenzhen 518055*, China*

Received 16 September 2017/Revised 29 November 2017/Accepted 8 January 2018/Published online 19 June 2018

Citation Lei K, Zhang L Z, Liu Y, et al. An event summarizing algorithm based on the timeline relevance model in Sina Weibo. Sci China Inf Sci, 2018, 61(12): 129101, https://doi.org/10.1007/s11432-016-9333-4

Dear editor,

Depicting superior punctuality and originality, Weibo has become increasingly critical and influential in China for online information acquisition and sharing. However, very few research has studied Weibo to investigate event summarizing even though most of the published Weibos are event-driven. Besides, we observe that the existing methods are unsuitable for processing tweets (short-texts with no obvious contextual relationships) although extensive researches have successfully extracted summaries from single-long-documents [1–5]. Thus, effectively summarizing the antecedents and consequences of events from massive tweets is still considered to be challenging. In this study, a timeline relevance model was initially established to estimate the popular period of the event. Further, a summarizing algorithm was designed to mine and summarize the related events. Finally, the experimental results depict the superiority of our summarizing algorithm. Our study can help users to quickly identify and grasp the event that is related to the topic in a limited time frame. Furthermore, these methods can also be applied to other situations, such as multiple short-text summarization and event detection.

Timeline relevance model. To estimate the start and end times of topic-related events, this research constructed a timeline relevance model based on the temporal correlation of hot topic queries and tweeting behaviors on Weibo, which can significantly improve the comprehensiveness of the retrieval and is critical for detecting related events.

According to the analysis report regarding the topics of Sina Weibo, the lifecycle of a topic usually consists of four stages, which are the birth, growth, mature, and recession. When the popular period of an event is to be estimated, the following three aspects should be considered: (1) the initiation of the birth period of an event; (2) the duration of its growth and its mature periods; and (3) the termination of its recession stage. Thus, the features of these stages in the lifecycle of hot topics are examined by considering the timeline of the searching activity for each topic to be the baseline, and the hourly gap between the duration of the searching and tweeting activities during different stages are compared. Table 1 depicts the results, where the hourly gap is expressed using its absolute value.

Based on the statistical results, a timeline relevance model was further established to estimate the start and end times of the related event. Specifically, we let s_{grow} denotes the start time of the growth period (when the hashtag on the popular list emerges), t_{start} denotes the time when the topic enters the birth period (the start of the tweeting activity), s_{end} denotes the end time of the recession stage (when the hashtag disappears from

* Corresponding author (email: leik@pkusz.edu.cn)

Table 1 The hourly gap between the searching and tweeting activities during different stages

Different stages of the lifecycle	Hourly gap (h)	Proportion (%)
$\lvert t_{\text{start-search}}(\text{Growth period}) - t_{\text{start-tweet}}(\text{Birth period})\rvert$	12–24	76.31
$\lvert t_{\text{end-search}}(\text{Growth period}) - t_{\text{end-tweet}}(\text{Birth period})\rvert$	3–5	92.54
$\lvert t_{\text{end-search}}(\text{Growth period}) - t_{\text{end-tweet}}(\text{Birth period})\rvert$	6–8	74.63

the popular list), t_{end} denotes the end time of the tweeting activity, and the duration of the related event can be defined as $[t_{\text{start}}, t_{\text{end}}]$, where t_{start} and t_{end} can be derived as follows:

$$t_{\text{start}} = s_{\text{grow}} - 12h, \tag{1}$$

$$t_{\text{end}} = s_{\text{end}} + 6h. \tag{2}$$

Since it is easy to evaluate the values of s_{grow} and s_{end}, the period during which the tweets mentioning the topics emerged and disappeared was estimated using the timeline relevance model.

Event summarizing algorithm. Based on the timeline relevance model, the duration of the related event was defined as an exploring period to retrieve the topic-related tweets, and an event summarizing algorithm was designed to further explore the ins and outs of the given topic. Particularly, the complete algorithm includes a hashtag expanding language model, binaryKmeans_HAC clustering algorithm, and user authority based hybrid TF-IDF algorithm.

• Hashtag expanding language model. Most of the "initial queries" in this study were represented using a single hashtag that was short and had ambiguous context. Thus, this hashtag would result in a one-sided retrieval. To enhance the expression of the query and to associate more relevant tweets, a hashtag expanding language model with dynamic pseudo relevance feedback was proposed.

In this study, the "initial query" was expanded by considering the following two aspects: (1) apply the method proposed in our previous study that considers both the similarity of Weibo users and tweets to calculate the similarity between various hashtags [6]; (2) based on the observed pseudo relevance feedback, estimate the weight of the various terms in tweets that were retrieved by the "initial query", and further select the top k ranked terms to expand the "initial query". Finally, the weight of a term can be derived as follows:

$$W_{t_i} = \lambda \times W_{t_i}1 + (1-\lambda) \times W_{t_i}2. \tag{3}$$

W_{t_i} represents the correlation weight of the term t_i; $W_{t_i}1$ denotes the similarity between t_i and the hashtag (if t_i is another hashtag), which is derived by

$$W_{t_i}1 = \begin{cases} \text{Sim}(T, t_i), & \text{Sim}(T, t_i) \geqslant 0.11, \\ 0, & \text{else}, \end{cases} \tag{4}$$

$$\text{Sim}(T, t_i) = \frac{S(U_a, U_b)}{\max\{S(U_i, U_j), i \in H, j \in H\}} + \frac{S(T_a, T_b)}{\max\{S(T_i, T_j), i \in H, j \in H\}}, \tag{5}$$

$$S(U_a, U_b) = \frac{|U_a \bigcap U_b|}{|U_a \bigcup U_b|}, \tag{6}$$

$$S(T_a, T_b) = \frac{|T_a \bigcap T_b|}{|T_a \bigcup T_b|}, \tag{7}$$

where $S(U_a, U_b)$ represents the similarity of Weibo users who tweeted about the two corresponding hot topics, a and b; H denotes the set of hot topics; U_a and U_b represent the set of users who have tweeted about a and b, respectively. Similarly, $S(T_a, T_b)$ denotes the similarity of various tweets for the two topics, and T_a and T_b represent the set of tweets that mention either a and b.

$W_{t_i}2$ represents the term weight that calculated using the pseudo relevance feedback method [7]. First, the initial hashtag was used to retrieve the relevant tweets. Further, the TF-IDF value of each word in the related tweet set was calculated. Particularly, all the tweets were considered to be a big document when calculating TF. However, each tweet acts as an individual document while calculating IDF. Additionally, $W_{t_i}2$ was calculated using the formula that is given in

$$W_{t_i}2 = \text{TF}_{t_i} \times \text{IDF}_{t_i}. \tag{8}$$

Finally, these terms were ranked according to their W_{t_i} value, and then the terms with higher weight were selected to expand the "initial query".

• BinaryKmeans_HAC algorithm. To explore the different aspects of the topic-related event, a BinaryKmeans_HAC algorithm, which combines the bisecting K-means and agglomerative hierarchical clustering, was proposed to cluster the relevant tweets. First, the initial cluster was divided into two clusters. Further, the two clusters were repeatedly divided on the basis of higher SSE (sum of squared error) values into two clusters using the bisecting K-means algorithm until the number of clusters became equal to m. Additionally, we executed agglomerative hierarchical clustering to recursively aggregate the two most similar clusters into a bigger cluster, until the K clusters were observed to be sufficiently cohesive. Finally, the different clusters that were aggregated using Bina-

ryKmeans_HAC can be used to describe the various aspects of an event.

• User authority based hybrid TF-IDF algorithm. After expanding the query, retrieving and clustering the tweets, a user authority based hybrid TF-IDF algorithm was finally proposed to extract the summary from the tweet sets clustered by BinaryKmeans_HAC.

On one hand, the value obtained using TF-IDF was leveraged to depict the importance of a tweet by evaluating the importance of every term that it contained. Considering the particularity of tweets (short-texts but large in number), a hybrid TF-IDF algorithm was proposed to evaluate the weights of the tweets. For a topic, all the related tweets composed a complete document when calculating TF. Further, each tweet was treated as an individual document when calculating IDF. The algorithm can be represented as follows:

$$W(T_i) = \frac{\sum_0^{\mathrm{nf}(T_i)} w_{t_i}}{\mathrm{nf}(T_i)}, \tag{9}$$

$$w_{t_i} = \mathrm{TF}_{t_i} \times \mathrm{IDF}_{t_i}, \tag{10}$$

$$\mathrm{TF}_{t_i} = \frac{n_{t_i}}{\sum_k n_{t_k}}, \tag{11}$$

$$\mathrm{IDF}_{t_i} = \log \frac{|D|}{\sum_j d_j, t_i \in d_j}. \tag{12}$$

$W(T_i)$ indicates the weight of the tweet T_i, $\mathrm{nf}(T_i)$ represents the total number of the words in T_i, and w_{t_i} denotes the weight of the term t_i.

On the other hand, we observe that the participation of celebrities or authoritative users can significantly accelerate the dissemination of related events. Additionally, we observe that user authority can be reflected on the authority of the tweets to some extent. Therefore, the number of fans of the author was leveraged to represent the corresponding authority of the tweet. Combining the weight calculated using the hybrid TF-IDF algorithm, the author influence of the tweet was also considered into our calculation. Finally, the weight of a tweet can be defined as follows:

$$W'(T_i) = W(T_i) + \mathrm{Influence}(d_i), \tag{13}$$

$$\mathrm{Influence}(d_i) = \frac{|\mathrm{followers}|_{u_i}}{\max_{\forall u_j} |\mathrm{followers}|_{u_j}}. \tag{14}$$

Evaluation. To assess the performance of our enhanced event summarizing algorithm, both the proposed algorithm and the other four typical algorithms were applied to the Weibo dataset. Further, their experimental results were compared with the referenced summarization annotated by the artificial method. Finally, the precision, recall rate, and F1 value of the experimental results were calculated according to the ROUGE-1 criteria.

The results verify the superiority of our event summarizing algorithm. Both the timeline relevance model and hashtag expanding ensures the comprehensiveness of the retrieval. Additionally, the BinaryKmeans_HAC and user authority based hybrid TF-IDF algorithms ensure the diversity and accuracy of the summary. Therefore, the final precision and F1 value, especially the recall rate, have been improved using our proposed methods.

Application. The proposed study for detecting topic-related events and for extracting core information of events can save a lot of time that user have to spend for searching and retrieving. Furthermore, the proposed methodology can also be applied to other situations, such as event detecting, and short text summarization.

Acknowledgements This work was supported by Shenzhen Key Fundamental Research Projects (Grant Nos. JCYJ20170412150946024, JCYJ20170412151008290).

References

1 Yih W T, Goodman J, Vanderwende L, et al. Multi-document summarization by maximizing informative content-words. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, 2007. 1776–1782

2 Ferreira R, Cabral L D S, Freitas F, et al. A multi-document summarization system based on statistics and linguistic treatment. Expert Syst Appl, 2014, 41: 5780–5787

3 Li J X, Li L, Li T. Mssf: a multi-document summarization framework based on submodularity. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, 2011. 1247–1248

4 Wang D D, Li T, Zhu S H, et al. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, 2008. 307–314

5 Mihalcea R, Tarau P. TextRank: bringing order into texts. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, 2004. 404–411

6 Yu Q, Weng W T, Zhang K, et al. Hot topic analysis and content mining in social media. In: Proceedings of the 33rd IEEE International Performance Computing and Communications Conference (IPCCC), Austin, 2014

7 Chawla S, Bedi P. Query expansion using information scent. In: Proceedings of International Symposium on Information Technology, Kuala Lumpur, 2008