Postprint: Combined Model Based on Event Elements for Weibo Hotspot Event Summarization

Authors: Li Gang, Xu Wei, Wang Xinping

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] To assist readers in rapidly comprehending the context and evolution of events from massive Weibo reports generated by hot events, and to enhance the accuracy and readability of Weibo event summarization, this paper proposes a multi-model timeline summary extraction method for Weibo hot events based on event elements. [Method/Process] Tailored to the characteristics of Weibo text, the method extracts event summary keywords by integrating the topic model (LDA) with the mutual information maximum entropy model (MaRxEnt-MI), filters Weibo posts based on their dissemination value and thematic relevance, and generates timeline summaries in the format of time-summary keywords-summary Weibo posts. [Results/Conclusion] Evaluated on a manually annotated test set and compared with the traditional TextRank method, the proposed method achieves an 8%-13% improvement in F-score, with internal tests demonstrating significant enhancement in summary readability. The scale and event diversity of experimental texts and test sets require further expansion, and additional weighted strategy models should be considered to improve summarization accuracy. Experimental results and user feedback indicate that the proposed method effectively satisfies users' information needs for hot event summaries and enhances the accuracy of Weibo summary extraction.

Full Text

Preamble

Extracting Summaries of Hot Events on Microblogs Based on Event Elements Using a Combined Model

Li Gang, Xu Wei, Wang Xinping
School of Information Management, Wuhan University, Wuhan 430072

Abstract

[**Purpose/Significance**] To help readers quickly understand the context of hot events from massive microblog reports and improve the accuracy and readability of microblog event summaries, this paper proposes a multi-model timeline summary extraction method for hot microblog events based on event elements. [**Method/Process**] Addressing the characteristics of microblog text, we combine the strengths of the Latent Dirichlet Allocation (LDA) topic model and the Mutual Information Maximum Entropy (MaxEnt-MI) model to extract event summary keywords. We then screen microblogs based on their communication value and thematic relevance, generating timeline summaries in the format of time-summary keywords-summary microblogs. [**Result/Conclusion**] Using a manually annotated test set, our method achieves an 8%-13% improvement in F-score compared to the traditional TextRank method, with internal tests showing significantly improved summary readability. While the quantity and event richness of experimental texts and test sets need further expansion, and more weighting strategies should be considered to enhance accuracy, experimental results and test feedback demonstrate that our method effectively meets users' information needs for hot event summaries and improves the accuracy of microblog summary extraction.

With the popularization of the Internet and the rise of microblogs, various users—including ordinary netizens, online celebrities, news media, and government agencies—have adopted microblogs as their primary channel for obtaining news and publishing comments. When hot events occur, massive amounts of data accumulate on microblog platforms. However, due to characteristics such as severe colloquialism, short text length, serious semantic loss, abundant spam, and rapid information growth, readers struggle to quickly understand the full context of events. For hot events, the temporal dimension is a crucial component of event description, and extracting texts that represent the development of the event at important time points enables users to rapidly comprehend hot events.

In recent years, automatic microblog text summarization technology has gradually emerged. Related international research includes D. Inouye's proposal to use a combined TF-IDF algorithm for sentence scoring and ranking to generate multi-post microblog summaries by removing redundancy, as well as methods that first cluster microblogs and then extract important posts from each category as summaries. R. Swan and J. Allan manually designed event tables, extracting named entities for each time node and linking them chronologically as event timelines. Domestic research started later, with R. Long et al. using keyword graph clustering to select n microblogs most relevant to hot event content as summaries. X. Wan considered the influence of temporal factors in text structure and proposed the TimeTextRank algorithm for text summarization. However, most related research uses Twitter as the study object and rarely considers the importance of temporal factors. News media summarization research initially emphasized temporal elements, but similar research remains scarce in

microblog event summarization.

Through analysis of hot events and related news reports on microblogs, we found that event subjects, time, location, and persons are the event elements people care about most. From netizens' comments and concerns, the most frequent user inquiries relate to developments at important time nodes during event evolution. From the content of related microblogs, the distribution of event elements and entity words changes continuously across important time periods. Moreover, the overall data sparsity problem is significant, with large text volumes, severe colloquialism, and abundant useless information.

Therefore, this paper proposes using the LDA (Latent Dirichlet Allocation) model to extract topic keyword sets, then employing the maximum entropy mutual information model to address the disorderliness of keywords extracted by the topic model, optimizing the topic keyword list. By integrating event elements and a comprehensive measurement method of microblog influence to determine the importance of microblogs within important time points, we generate timeline summaries of hot events.

2 Research Approach and Related Models

2.1 Research Approach

Hot event timeline summary aims to output, at several important time periods of an event, a text set that can represent the development of the node event. These texts should comprehensively summarize the main content of online news reports about the event at that time node. The timeline summary process includes three main stages: important event feature extraction, timeline summary keyword extraction, and summary sentence output.

We first use the LDA model to extract hot topics and their keywords from microblog corpora, then combine event element extraction results, part-of-speech information, topic probabilities, and MaRxEnt-MI calculated word relationships to compute keyword weights and generate summary keywords. For microblogs within each time period, we calculate weights based on the previously generated summary keywords combined with the news value of the microblogs themselves, selecting high-weight sentences for summary generation, presented in the format of time-summary keywords-summary microblogs. The specific process is shown in [Figure 1: see original paper].

2.2 Microblog Hot Event Summarization

Compared with traditional document summarization, microblog event summarization exhibits different characteristics due to the nature of microblog products and information content. These characteristics inform the generation of summaries better adapted to microblog documents and users.

2.2.1 Characteristics of Hot Event-Related Microblog Content

(1) Microblog information belongs to short text, even shorter than other short texts. While some microblogs are severely colloquial and poorly structured at the sentence level, they demonstrate strong word expressiveness. Since a single microblog generally contains no more than four sentences, it is unsuitable to perform extractive summarization on individual microblog texts.

(2) Microblog information often focuses on one aspect of an event. With strong immediacy and interactivity, and due to text length limitations, microblog content typically concentrates on a specific time point and aspect of an issue. Unlike lengthy news reports that cover all aspects of an event, a single microblog cannot adequately summarize the event situation.

(3) High-quality microblog texts reporting on events contain relatively complete event elements, especially temporal elements. Whether from authoritative accounts such as media, government agencies, or other users, high-quality microblogs often include complete event elements and highlight temporal phrases due to word count limitations and communication needs. Complete event elements and prominent temporal phrases also improve the readability of summaries.

2.2.2 Definition of Microblog Hot Event Summaries Current research on microblog event summarization is limited. B. Sharifi, M. A. Hutton, and J. Kalita, referencing the hot event summary function of the WhatTheTrend system on Twitter, defined microblog event summarization as extracting microblogs most closely related to an event from all microblogs about that event. Like that study, microblog texts are too short for single-microblog summary extraction, so we select entire microblogs as part of the summary. However, B. Sharifi et al. did not consider the characteristic that single microblogs cannot adequately summarize hot events. Therefore, this paper selects a collection of multiple microblogs as the summary text while dividing the microblog collection by time periods. Thus, we define microblog hot event summarization as selecting the microblog collection that best summarizes the event situation from all microblogs within a certain time period.

2.2.3 Behavioral Characteristics of Microblog Users Participating in Hot Events

(1) The quality of event reporting on microblogs correlates with the publisher's identity. While news reports are written by journalists and published by news organizations with relatively consistent quality, microblogs lack an audit process, resulting in uneven content quality. However, because microblog users have followers or are social accounts of government agencies, media groups, websites, or celebrities, they bear social responsibility or must meet fan expectations when publishing information. Their microblogs are of high quality and provide accurate descriptions of certain

event aspects, making them ideal for summaries. Additionally, event witnesses or participants may generate high-quality, timely microblogs that attract numerous reposts, also making them ideal summary candidates.

(2) The frequency with which netizens view microblog information and participate in event discussions follows a daily periodic pattern. According to the "2015 China Social Media User Behavior Research Report," 47.5% of users check microblogs daily. Users generally pay attention to what information is available today and yesterday for hot events they follow. Therefore, organizing microblog summary sets by day aligns with user habits.

(3) Netizens' attention to event-related reports is not limited to one aspect, and they continue to follow events over time. While news coverage of events is aggregated, microblog coverage often focuses on individual aspects. User concerns vary, and they often follow up continuously. For example, during the February 6 Kaohsiung earthquake in Taiwan, microblog reports and user concerns on that day included multiple aspects such as earthquake location, epicenter, and casualties. On February 7, attention shifted to rescue efforts and casualty number updates. When browsing microblogs, users gather information from various aspects, but their information needs are often difficult to satisfy due to the large volume of microblog data.

In summary, for all microblogs related to a designated hot event, we divide them into multiple microblog collections by day. We then use the composite model proposed in this paper to extract summary keywords and filter out microblogs with high news value from each day's collection to form a summary microblog set. Since the temporal element is an important component of each microblog, we extract temporal elements from each microblog in the summary set and standardize them using the algorithm proposed in this paper. Microblogs containing temporal expressions in the summary set are sorted and output according to their extracted temporal expressions, while those without temporal expressions are placed at the end.

2.3 LDA Model Introduction

The basic principle of the LDA model is a three-layer Bayesian model that can model implicit topics in texts. Compared with traditional similarity calculation methods, LDA can automatically generate semantic topics from massive text data in an unsupervised manner. The LDA model considers documents as mixtures of topics, mapping high-dimensional text corpora to low-dimensional potential semantic space. It treats topics as distributions over word space to obtain relationships between texts, describing the document generation process. The model representation is shown in [Figure 2: see original paper].

In the LDA generation process, the joint distribution of observed and hidden variables is calculated as in formula (1):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod p(\beta) \prod p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:k}, z_{d,n}) p(\theta_d) \quad (1)$$

where β represents topics, θ represents topic probabilities, z represents the topic of a specific document or word, and w represents words. $\beta_{1:K}$ is the complete topic set, where β_K is the word distribution of the k-th topic (see [Figure 2: see original paper]). The proportion of this topic in the d-th document is θ_d, where $\theta_{d,k}$ represents the proportion of the k-th topic in the d-th document (see [Figure 2: see original paper]). The complete topics of the d-th document are Z_d, where $z_{d,n}$ is the topic of the n-th word in the d-th document (gray circle in [Figure 2: see original paper]). All words in the d-th document are denoted as w_d, where $w_{d,n}$ is the n-th word in the d-th document, with each word being an element of a fixed vocabulary. $p(\beta)$ represents selecting a specific topic from the topic set, $p(\theta_d)$ represents the probability of that topic in a specific document, $p(z_{d,n}|\theta_d)$ is the topic of the n-th word in the document given the topic, and $p(w_{d,n}|\beta_{1:k}, z_{d,n})$ is the joint distribution of the topic of the n-th word in the document and the word itself. The product calculates the dependency of random variables.

Posterior distribution calculation is shown in formula (2):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \quad (2)$$

In practice, for the numerator, it is easy to count given the corpus. The denominator's computational load becomes impossible to calculate directly as text volume increases. If the corpus comprehensive word bank exceeds one million, containing n words, each word has m observation combinations, and the prior probability must be obtained through accumulation, making the computation enormous. Therefore, an approximate solution method is needed.

Common posterior distribution methods include Expectation Propagation, Laplace approximation, and Gibbs sampling. This paper uses the Gibbs sampling method to estimate the posterior distribution of current feature words and topics. The Gibbs sampling algorithm process is: first, sample initial topics for all words in the corpus; second, scan the corpus and recalculate topics for each word using the Gibbs sampling formula, updating them in the corpus; third, repeat the resampling process until convergence; finally, obtain the topic-word probability matrix to ultimately acquire the probability of each feature word under multiple topics in the document.

The LDA model has many successful applications in summary extraction. R. Arora and B. Ravindran used the LDA model to calculate word weights under each topic, obtaining sentence word weight vectors, and used singular value decomposition to extract sentences that best represent topic meanings as document summaries. The hybrid LDA and SVD model effectively reduced repetitive and redundant parts in summaries. Y. Petinot, K. Mckeown, and K. Thadani

proposed the hLLDA model, establishing correspondences between each label and topic, then extracting summaries through hierarchical categories.

2.4 MaxEnt-MI Model

In extracting microblog summary keywords, we generally examine the internal tightness and external boundaries between keywords. Higher internal tightness indicates better phrase integrity, meaning close internal word string connections. External boundaries can measure the independence of the phrase's overall expression; higher indices indicate stronger semantic functional expression of the multi-word phrase, such as "发生地震" (earthquake occurrence) and "危机公关" (crisis PR). Common internal methods include t-score, mutual information, and log-likelihood values. Common external methods include left-right entropy.

Based on internal tightness measurement, we select the word mutual information model; for external boundary measurement, we select the maximum entropy model. Combining these two models, we propose the MaxEnt-MI model to measure the degree of connection between keywords. This connection degree (defined as MEMI) indicates stronger semantic functionality and higher phrase integrity. Higher MEMI values suggest stronger semantic function and higher phrase completeness. The process of using this model to process corpora is shown in [Figure 3: see original paper].

2.4.1 Word Mutual Information Calculation Mutual information reflects the degree of interdependence between two words. Based on research by Sun Maosong and Luo Shengfen et al. on the effectiveness of various statistical measures in Chinese word extraction tasks, mutual information demonstrates the best word extraction performance, with strong complementarity between multiple methods. Building on this research, we select mutual information as the statistical measure to determine the internal combination correlation degree between word strings. The calculation method is shown in formula (3):

$$MI(X,Y) = \log_2 \frac{P(X,Y)}{P(X)P(Y)} \tag{3}$$

where $P(X)$ represents the probability of word X appearing, and $P(X,Y)$ represents the probability of words X and Y appearing in all second-order word strings. For example, if "台湾 → 地震" (Taiwan→earthquake) appears 314 times in the corpus and there are 2000 total second-order word strings, then $P(X,Y) = 314/2000$. By definition, higher mutual information means tighter intrinsic binding between two words; conversely, there may be phrase boundaries between the two words.

2.4.2 Left-Right Information Entropy Calculation Generally, entropy measures the uncertainty of random variables. Assuming random variable X

can take a limited number of values with calculable probabilities $P(X_i)$, the entropy of X is defined as $H(X)$, calculated as:

$$H(X) = -\sum_{x_i \in X} P(x_i) \cdot \log_2 P(x_i) \tag{4}$$

This paper applies the definition of information entropy to measure the entropy of left and right boundaries in multi-word phrase expressions, thereby measuring the external boundary of multi-word phrases. The formulas are:

$$\begin{aligned} FL(W) &= -\sum p(aW|W) \cdot \log_2 P(aW|W) \\ FR(W) &= -\sum p(Wb|W) \cdot \log_2 P(Wb|W) \end{aligned} \tag{5}$$

where $FL(W)$ and $FR(W)$ represent the left and right entropy of the target word string combination, W represents all word strings, A represents the set of words appearing to the left of the target word string, and a represents a specific word in that set. B represents the set of words appearing to the right of the target word string, and b represents a specific word. $P(aW|W)$ represents the probability of word a appearing to the left of the target word string, and $P(Wb|W)$ represents the probability of word b appearing to the right.

2.4.3 Calculating Word Connection Degree We define the probability that two keywords can form a phrase with strong semantic functionality and high internal connection degree (MEMI) as the weighted sum of the above three statistics:

$$S(W) = \alpha \cdot MI(W) + \beta \cdot FL + \gamma \cdot FR \tag{6}$$

This part mainly sorts keyword groups based on obtained mutual information values and left-right information entropy. Weight selection is based on test corpus calculation results and manual evaluation effects, with fine-tuning to obtain estimated values. After repeated debugging, parameters $\alpha = 1$, $\beta = 0.5$, and $\gamma = 0.6$ yield good results. Testing shows that results sorted by MEMI values indicate each word only has strong connections with certain words, which are often fixed collocations, verb-object combinations, or modification relationships with strong semantic functionality.

2.5 Combination and Improvement of LDA and MaxEnt-MI Models

When processing microblog corpora, the LDA model effectively solves text sparsity problems, and its unsupervised nature reduces dependence on domain knowledge when processing hot microblog events. However, the LDA model has several issues requiring model improvement or combination:

(1) The LDA model's fundamental assumption when processing corpora is that the order of topics and words in documents is irrelevant—that is, word items are exchangeable. Therefore, keywords extracted under topics through LDA form a disordered combination. However, keyword order is crucial for events, and word sequence also affects summary readability.

(2) LDA yields the probability that each word under a topic belongs to that topic but does not involve the importance of words under the topic.

(3) The LDA model's topic distribution skews toward high-frequency words, causing many topic-representative words to be overwhelmed by high-frequency words and reducing the model's topic expression capability.

(4) The LDA model requires setting the number of topics, and different topic numbers yield different extraction effects.

The MaxEnt-MI model processes microblog corpora with low time complexity and high efficiency, revealing word relationships such as order, fixed collocations, verb-object combinations, and modification relationships under microblog sparse corpus conditions. It demonstrates strong semantic functionality and easy comprehension. However, this model only considers features between words, ignoring features between words and microblogs, word position relationships, and amplifying the function of low-frequency words in corpora, causing words with poor expression capability to be selected as text representations.

Nevertheless, MaxEnt-MI and LDA models are well-suited for microblog short text summarization and exhibit strong complementarity. MaxEnt-MI solves LDA's word ordering problem and improves keyword readability. After LDA extracts topics and generates keywords, the clustering facilitates further MaxEnt-MI processing. Following LDA's topic-keyword acquisition, the MaxEnt-MI model can complete phrase recognition based on keyword part-of-speech and weight features, combined with event element information to complete event keyword screening and obtain temporal summary keyword identification.

In addition to the combined model approach, this paper performs a series of auxiliary operations to improve model accuracy: (1) Through word segmentation and part-of-speech tagging of all corpora, we extract specified content word lists, identify relevant person, location, and organization names using named entity recognition, extract microblog keywords using the TextRank method, and combine these three word lists after deduplication to form our event topic word list. This filters out semantically weak words (e.g., stop words, conjunctions), increasing the proportion of strongly expressive words in LDA sampling for better results. (2) We determine the number of important event time nodes to set the number of topics extracted by the LDA model, making LDA-extracted topics more accurate. (3) We use the summary keywords output by MaxEnt-MI for readability optimization.

3 Experimental Process

3.1 Preprocessing

Preprocessing mainly involves segmenting microblogs related to hot events, removing stop words, eliminating microblog special symbols (emoticons, topic symbols, URLs, @nickname forwarding/replies), named entity recognition, and part-of-speech tagging. This paper uses the open-source ICTCLAS segmentation system to complete word segmentation, part-of-speech tagging, and named entity recognition, which supports user-defined dictionaries with fast segmentation speed and high accuracy.

3.2 Feature Extraction

Feature extraction involves three tasks: obtaining event-related topic word lists, extracting important event time nodes, and extracting event elements.

3.2.1 Obtaining Event Topic Word Lists Event topic words consist of all content words in the event microblog collection. The extraction process is: (1) Perform word segmentation and part-of-speech tagging on all corpora to extract content word lists with parts of speech as nouns, verbs, adjectives, and adverbs. (2) Use ICTCLAS named entity recognition to identify relevant person, location, and organization names. (3) Use the TextRank method to extract keywords from each microblog. Following R. Long et al.'s approach, we screen microblogs with length greater than 10 after processing. For microblog sentences, we use formula (7) to determine the number of keywords to extract:

$$N = \max\left(\left\lfloor \frac{L}{\beta} \right\rfloor, \alpha\right) \tag{7}$$

where L represents sentence length, β is the compression ratio, N is the number of noun content words in the sentence, and α is the minimum number of keywords per sentence. Generally, longer sentences with more noun content words contain more keywords. Through experimentation, $\beta = 5$ and $\alpha = 2$ yield the best results.

Combining these three word lists after deduplication forms our event topic word list, which filters out semantically weak words (e.g., stop words, conjunctions).

3.2.2 Obtaining Important Time Node Expressions After hot events occur, many microblog reports contain temporal phrases, which are particularly important for event summary extraction and representation, serving as crucial information for showing the temporal context. However, expressions are relatively messy and require unified processing and logical calculation: (1) Use ICTCLAS tools to identify temporal named entity expressions such as "2004 年 3 月 3 日" (March 3, 2004), "昨日" (yesterday), and "以前" (before), dividing results into standardizable and non-standardizable categories. (2) Standardize

standardizable times using custom rules, then perform temporal calculation. Text temporal calculation proceeds in two steps: first, obtain two elements for temporal calculation (time benchmark and temporal relationship phrase). If the text contains temporal phrase relationships (e.g., "三天以后" - three days later, "昨天下午" - yesterday afternoon), search for temporal benchmarks before the phrase (e.g., "今天" - today, complete or incomplete time expressions). If the text contains temporal relationships but no temporal benchmark, use the microblog event as the benchmark. Second, calculate time based on the temporal benchmark and relationship phrase, converting original temporal relationship phrases to standard times. (3) Count standardized times extracted from the text to obtain the important time node collection.

Using all microblogs about the "Kaohsiung earthquake" event (from February 6, 2016 to February 15, 2016) as the dataset, temporal phrase statistics and standardization examples are shown in and .

3.2.3 Event Element Identification Event elements are important components of hot events. This paper extracts event elements according to the 5W1H framework from journalism: where (location), when (time), who and whom (participants), what (specific action), and how (result). For named entities such as location, person, and organization, we directly use ICTCLAS for annotation. Temporal elements are processed using the temporal phrase identification method described above. The "what" element (specific action) is extracted through syntactic analysis and predefined rule templates (NP1+V+NP2, NP+V, V+NP, V+Va, etc.) to extract subject-verb-object triples. The "how" element extracts words with strong emotional tendencies from statements using an emotion dictionary.

Using the above methods, we can obtain the event element set {T, L, NT, P, V, S} from a single microblog, where T represents the temporal element set, L represents location elements, NT represents organization/location elements, P represents person/title words, V represents action words, and S represents emotional tendency words. When processing documents for a certain time period or all documents, we simply integrate the elements from each microblog.

After obtaining event element sets on time periods and the entire corpus, we can calculate the weight of each element in the time period. The weight quantification formula is:

$$WF(f_i) = \frac{tf(f_i + T)}{sum(T)} \tag{8}$$

where $WF(f_i)$ represents the weight of word f_i in event element F within the element set in time period T, $tf(f_i + T)$ represents the frequency of word f_i in time period T, and $sum(T)$ is the total number of elements in event element F in time period T.

3.3 Combined Model for Extracting Event Summary Keywords

3.3.1 Extracting Topic Keywords We use the event topic word list extracted in the previous step to filter microblog corpora, selecting processed microblogs with length greater than 10 to train the LDA model using Gibbs sampling: (1) Randomly assign a topic to each word in each document of the document set. (2) Scan the document set and resample topics for each word using the Gibbs sampling formula, updating them in the document set. (3) Repeat the resampling process until convergence. (4) Obtain the topic-word probability matrix.

3.3.2 Keyword Weight Assignment For extracted topic keyword lists, we need to calculate keyword weights considering two important attributes: (1) The descriptive capability of keywords for events generally relates to part-of-speech and the role they play as event elements. For example, "地震" (earthquake) is a highly expressive word. (2) The keyword appears frequently in the selected time period but not in other periods, measuring word volatility across time spans using standard deviation. Combining these factors, we calculate keyword weight W_i using:

$$W_i = \alpha \cdot \frac{\sum_{k=1}^{T/\Delta T} \left(F_k - \frac{F_i}{T/\Delta T}\right)^2}{F_i} \tag{9}$$

where T refers to the time span of hot event microblogs, ΔT is the selected time interval, F_k is the word frequency in the corresponding time period, F_i is the total word frequency, and α corresponds to the event element of the word. Through data preprocessing and observing keyword output effects under various parameter values, we finally determine location = 0.75, person name = 1, organization = 0.5, time = 0.75, and other words = 0.2.

Using the "Kaohsiung earthquake" event as an example with data from February 6, 2016, LDA topic keyword results are shown in .

3.4 Using MaxEnt-MI Model to Merge Keywords and Generate Summary Keywords

Based on obtained keywords and their weights, we select keywords with $W_i > I$. According to word relationship pairs extracted by MaxEnt-MI (using the Taiwan Kaohsiung earthquake event as an example, partial results shown in), we divide keywords into two sets: event object set $Subj\{Pair_1, ..., \}$ and event action set $Action\{Pair_1, ..., \}$. The construction rules are: (1) $Subj - Pair_i$ selects keyword combinations as n+n, a+n (where n is selected according to detailed part-of-speech tagging as nt, *nr, nn*), and adds named entity nouns. (2) $Action - Pair_i$ selects keyword combinations as n+v, v+n, v+v, v+t, adv+v, a+v, and adds verb content words. The final output is event keyword summaries in the form of $Subj + Action$ word group combinations.

Using the Kaohsiung earthquake event as an example with data from February 6, 2016, MaxEnt-MI model extracted relationship word pair results and event keyword summary examples are shown in .

3.5 Screening Microblogs as Event Timeline Summaries Based on Summary Keywords

After obtaining microblog event summary keywords, we need to output sentence sets that can represent event development within each time period, reflecting the main content of online news reports about the event at that time node. We extract important microblog sentences as summaries for each time node. Unlike news summarization, the medium differs but the goal is the same. We can refer to news value definitions to establish sentence screening criteria. News value refers to the sum of factors in news that satisfy public demand, or the sum of social value. The five elements of news value in communication studies include timeliness, importance, prominence, proximity, and interest. According to the factual principle of news timeliness, summary sentences should contain event elements (time, person, location, action). Good news summary texts contain more event elements. Prominence considers the publisher's influence and whether they are key dissemination nodes (big V, highly reposted microblogs). Proximity considers how microblogs containing event keywords reflect relevance to the event. Therefore, implementation requires comprehensive consideration of these elements. Combining journalism's definition of news value, we propose the following formula to screen summary microblogs:

$$P(i) = \frac{(4E_T + 2E_P + 2E_L + 2E_V)}{\max(4E_T + 2E_P + 2E_L + 2E_V)} \cdot \frac{K_i}{K_t} \cdot H_i \tag{10}$$

The first bracket calculates whether the microblog contains event elements. E_T indicates whether a standardized temporal phrase is included (1 if yes, 0 if no). E_P indicates whether a person name is included, E_L indicates whether a location name is included, and E_V indicates whether verb content words are included. If a microblog contains all elements, it is a possible news summary from the perspective of news expression.

The second bracket compares the microblog's keywords with our combined model's extracted event summary keywords. K_i represents the number of keywords hit by the microblog, and K_t represents the total number of keywords in the topic. If a microblog overlaps more with the topic's keywords, it has greater relevance to the topic.

The final H_i represents the microblog's social value. From the perspective of microblog news value, we consider H_i related to the blogger's type, microblog popularity (comments, reposts, likes), and blogger influence. For example, celebrities' microblogs about hot events may not meet microblog news value standards in content, but their participation itself is part of the hot event, and many netizens hope to see such microblogs when understanding event developments.

Additionally, government and official media reports on hot events may interest users more than those from microblog influencers and ordinary users. We propose a simplified microblog social value measurement method: (1) Determine user category based on crawled user account tags (enterprise, media, government, celebrity, website, group, campus, influencer, ordinary user, other). For high-influence types (celebrity, government, media, enterprise, website), directly output $H_i = 1$. (2) For medium-influence users (group, campus, influencer) with repost numbers below half of all event microblogs, $H_i = 0.6$; above half, $H_i = 1$. (3) For ordinary users and others with repost numbers below half, $H_i = 0.4$; above half, $H_i = 1$.

Using the Kaohsiung earthquake event as an example with data from February 6, 2016, partial extracted microblog summaries are shown in .

4 Experimental Results Analysis

We selected representative events from the "2016 Annual Social Hot Events Network Public Opinion Report"—the "Taiwan Kaohsiung 6.7 magnitude earthquake" and the "Heyi Hotel female assault"—using event titles and extended query words as search keywords. We crawled approximately 20,000 microblogs as experimental corpora. Microblog collection samples and conditions are shown in .

Since summary extraction differs from general NLP tasks, with the difficulty that standard answers are not unique, automatic evaluation is challenging. Many scholars have proposed automatic evaluation methods based on text characteristics, but these are complex to implement and beyond this paper's scope. Most evaluations use manual internal evaluation to generate standard summaries, which we also employ. Due to high requirements for evaluators, insufficient literary accomplishment may affect results. We adopt a strategy: since we extract microblog sentences representing hot event development status from microblog collections, manual evaluation only requires selecting microblogs that can serve as summaries. Due to workload, we only use two events ("Kaohsiung 6.7 magnitude earthquake" and "Heyi Hotel female assault") as evaluation data. To reduce the impact of individual literary accomplishment, we use Baidu Baike entries on hot events as references, which are compiled by numerous netizens and have relatively high accuracy.

After manual evaluation, we obtain manually extracted summary collections as test sets. We use the common F-measure as the evaluation standard:

$$F\text{-measure} = \frac{2 \cdot R \cdot P}{R + P} \tag{11}$$

where N is the number of summaries extracted by our method that match manual summaries, N_p is the total number of manually annotated summaries, and N_r is the total number of summaries extracted by our method.

For comparison, we apply the TextRank algorithm from news summarization to microblog corpora, treating microblogs within a certain time period as news texts to extract microblog summaries. Results are shown in .

Experimental results show that our method's F-value is 8%-13% higher than the TextRank method, proving that our method improves the quality of timeline summaries for microblog hot events. Testers also reported that keywords plus microblog sentence summaries effectively show the context of events on the timeline, which is important for helping users understand hot events and monitor events.

Research limitations include high recall but limited accuracy improvement in extracted microblog summaries. Some microblogs with low popularity but high summary value cannot be extracted. These microblogs are often not hot but reflect important event progress, such as "Members of the civilian rescue organization Gongyang Team from Zhejiang have arrived at Tainan Weiguan Building and been permitted to enter the collapsed building area to participate in rescue." Additionally, there is much room for optimization in key processes such as keyword selection, requiring further research.

In summary, building on traditional automatic summarization and news summarization research, combined with journalism communication characteristics and using the timeline as an organizational clue, we implement timeline summary extraction for hot events in the form of keywords plus key microblogs. Experimental results show that our method significantly improves recall compared to TextRank. We innovatively propose combining the LDA model and mutual information maximum entropy model to improve summary keyword accuracy and readability, ultimately obtaining timeline event summaries that meet user information needs.

References

[1] GOLDSTEIN J, KANTROWITZ M, MITTAL V, et al. Summarizing text documents: sentence selection and evaluation metrics [C]// SIGIR 2007: proceedings of the international ACM SIGIR conference on research and development in information retrieval. Amsterdam: DBLP, 2007: 867-868.

[2] CANHASI E, KONONENKO I. Multi-document summarization via Archetypal Analysis of the content-graph joint model [J]. Knowledge and information systems, 2014, 41(3): 821-842.

[3] CAI X, LI W. Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization [J]. IEEE transactions on audio, speech & language processing, 2012, 20(5): 1597-1607.

[4] 王红玲, 张明慧, 周国栋. 主题信息的中文多文档自动文摘系统 [J]. 计算机工程与应用, 2012, 48(25): 132-136.

[5] LUO Y, XIONG S. A combination scheme for distributed multi-document

summarization [J]. Journal of intelligence, 2013, 64(1): 94-102.

[6] INOUYE D. Multiple post microblog summarization [J]. Reuresearch final report, 2010(1): 34-40.

[7] SWAN R, ALLAN J. Automatic generation of overview timelines [C]// International ACM SIGIR conference on research and development in information retrieval. Athens: DBLP, 2000: 49-56.

[8] LONG R, WANG H, CHEN Y, et al. Towards effective event detection, tracking and summarization on microblog data [C]// International conference on web-age information management. Berlin: Springer-verlag, 2011: 652-663.

[9] WAN X. Timed TextRank: adding the temporal dimension to multi-document summarization [C]// Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 1999: 121-128.

[10] SHARIFI B, HUTTON M A, KALITA J. Summarizing microblogs automatically [C]// Human language technologies: the 2010 conference of the North American chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 685-688.

[11] GAGLIO S, LORE G, MORANA A M. Real-time detection of twitter social events from the user's perspective [C]// IEEE international conference on communications. London: IEEE, 2015: 1207-1212.

[12] WANG Y. Distributed Gibbs Sampling of Latent Topic Models: The Gritty Details [EB/OL]. [2017-04-10]. http://www.52ml.net/wp-content/uploads/2014/04/LDA-wangyi.pdf.

[13] CNNIC. 2015 China Social Media User Behavior Research Report [R/OL]. [2016-02-11]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/sqbg/201604/P020160722551429454480.pdf, 26.

[14] PORTEOUS I, NEWMAN D, IHLER A, et al. Fast collapsed Gibbs sampling for latent dirichlet allocation [C]// ACM SIGKDD international conference on knowledge discovery and data mining. Las Vegas: DBLP, 2008: 569-577.

[15] ARORA R, RAVINDRAN B. Latent dirichlet allocation and singular value decomposition based multi-document summarization [C]// IEEE international conference on data mining. Pisa: DBLP, 2008: 713-720.

[16] PETINOT Y, MCKEOWN K, THADANI K. A hierarchical model of web summaries [C]// The meeting of the Association for Computational Linguistics: human language technologies, proceedings of the conference. Oregon: DBLP, 2012: 670-675.

[17] 范小丽, 刘晓霞. 文本分类中互信息特征选择方法的研究 [J]. 计算机工程与应用, 2010, 46(34): 123-125.

[18] SHANNON C E, WEAVER W. The mathematical theory of communication [J]. Physics today, 1962: 97-117.

[19] 张小平, 周雪忠, 黄厚宽, 等. 一种改进的 LDA 主题模型 [J]. 北京交通大学学报, 2010, 34(2): 111-114.

[20] 张华平. NLPIR 汉语分词系统 [EB/OL]. [2014-01-15]. http://ictclas.nlpir.org.

[21] ZHU H D, ZHAO X H, ZHONG Y. Feature selection method combined optimized document frequency with improved RBF network [C]// Advanced data mining and applications, international conference, Adma 2009. Beijing: DBLP, 2009: 796-803.

[22] 何玲, 胡小强, 袁玖根. 麦克卢汉媒体观下微媒体的 5W 分析 [J]. 传媒, 2013(12): 55-57.

[23] 杨保军. 论新闻价值关系的构成 [J]. 国际新闻界, 2002(2): 55-57.

[24] 郝雨. 回归本义的“新闻价值”研究 [J]. 上海大学学报社会科学版, 2006, 13(6): 69-74.

Author Contributions

Li Gang: Responsible for research direction and paper revision guidance.
Xu Wei: Responsible for model design, data analysis, main content writing, and paper revision.
Wang Xinping: Responsible for paper proofreading and data analysis.

Note: Figure translations are in progress. See original paper for figures.