Knowledge-Based Systems 164 (2019) 107–121

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Social media based event summarization by user–text–image co-clustering

Xueming Qian *, Mingdi Li, Yayun Ren, Shuhui Jiang

Department of Information and Communication Engineering, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
Key Laboratory for Intelligent Networks and Network Security, Ministry of Education, Xi'an Jiaotong University, Xi'an, China

HIGHLIGHTS

- We proposed a coarse-to-fine filtering method to eliminate the irrelevance.
- We proposed a user-cluster based subevent determination approach.
- We proposed a user based event summarization approach and ensured its performance.

ARTICLE INFO

Article history:
Received 5 February 2018
Received in revised form 29 August 2018
Accepted 17 October 2018
Available online 3 November 2018

Keywords:
Event summarization
Social media
Image summarization
Co-clustering

ABSTRACT

Microblogging services have changed the way that people exchange information. There will generate a large number of data on the web once popular events or emergencies occur, including textual descriptions about the time, location and details for the event. Meanwhile users can review, comment, spread the event conveniently. It has always been a hot issue that how to use this mass of data to detect and predict breaking events. While existing approaches mostly only focus on event detection, event location estimation and text-based summary, a small amount of works have focused on event summarization. In this paper, we put forward a new social media based event summarization framework, which comprises of three stages: (1) A coarse-to-fine filtering model is exploited to eliminate irrelevant information. (2) A novel User–Text–Image Co-clustering (UTICC) is proposed to jointly discover subevents from microblogs of multiple media types—user, text, and image. (3) A multimedia event summarization process is designed to identify both representative texts and images, which are further aggregated to form a holistic visualized summary for the events. We conduct extensive experiments on Weibo dataset to demonstrate the superiority of the proposed framework compared to the state-of-the-art approaches.

1. Introduction

RECENT years we have witnessed that the development of social networks services changes the way in which people live, work and communicate. Today there are many social media services such as Facebook, Twitter, Yelp, Wechat, Weibo and etc., which benefit us to acquire and spread information. Especially with the boom of smartphones, we can access the Internet conveniently at any time and any place. Through the smartphone and the social media services, we can share what happened in our surroundings by means of texts, images and videos. Meanwhile, users on the web can view and comment on the content shared by world-wide users. Social media and smartphone promote each other by their prosperity. For instance, Weibo, as one of the most popular social networking platforms in China, is a text and image platform of information distribution and sharing. It has attracted more than 212 million active users, and the number of messages posted daily has reached 100 million by the end of 2015 [1,2]. Users are allowed to share multimedia content on such platforms including texts, images and video links.

An example of microblog in Weibo is shown in Fig. 1. We present what different names refer to in it, e.g., text, image, and user information and other useful data denoted by **else** such as re-post number. For instance, we call the textual information in each microblog as **text**, image information as **image**, user name, real name, hometown, gender etc. as **user information**. Each microblog mainly contains text, user information, comment number, repost number, attitude number and posting time. In addition, many of them contain image, which is very convenient for us to perform visualized summary for events.

With the wide availability of information sources, rapid information propagation and ease of use, Weibo has played a very important role in social events detection. In recent years, many works aim to detect hot events and breaking events by using hashtags, trending topics and common messages [3–19]. Many

* Corresponding author.
E-mail address: qianxm@mail.xjtu.edu.cn (X. Qian).

https://doi.org/10.1016/j.knosys.2018.10.028
0950-7051/

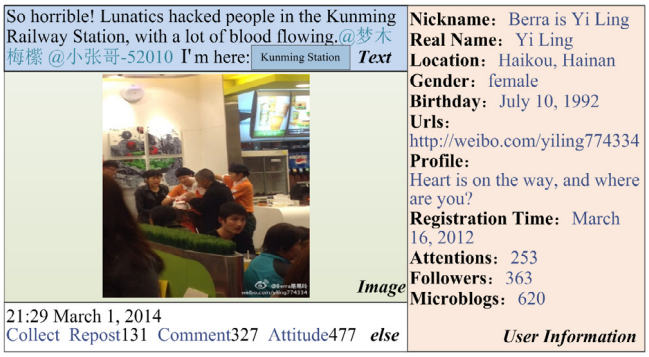

Fig. 1. An example of microblog in Weibo. Each microblog includes text and user information, etc. And many of them contain images.

works are proposed to estimate the location where the image are taken by a large geo-tagged image set [20–22]. There are also many works that aim to text-based event summarization [13,23,24]. With the rapid advancement of online social media platform, it brings us various media types such as images and video links. Since images typically convey a much more comprehensive impression of a specific situation compared to the limited text content of a microblog, we have to both take text and image into consideration for multimedia summarization of events. It is a key step for event understanding and cognition. In addition, there is also much user information as shown in Fig. 1, e.g., user name, id, hometown, followers' count, gender particularly in Weibo, which bring us much valuable information to analyze microblog popularity posted by different users. Users with similar interests may concern similar events, and post similar texts and images. Based on this, we introduce user information in our event summarization framework to enhance the relevance and diversity of textual and visual summary.

Multimedia based event summarization is important and meaningful, while it has several challenges: (1) Different users have different means of expression and writing habits in blogging. They may use different words to describe the same event. (2) The social media data generated by users contain a lot of noises. For example, images are not tagged, and there exist semantic gap between text and image. Also, some images are with poor illumination, especially for the events happened in the evening. It is very challenging to summarize the event from different viewpoints only relying on visual or textual features, which we call subevents in the following part of this paper. (3) Different user has different contribution in event summarization. For example, some users may choose irrelevant images for microblogs. Especially many salesmen utilize the microblog platform to broadcast their products. Thus there are many product photos in their shared photos, which are irrelevant to the events.

In order to solve the above challenges, some works carry out textual summarization and textual visual summarization from multiple source data, such as Weibo, Twitter [1,2,25–30]. We propose coarse-to-fine filtering method to reserve the relevant texts and images, and remove the irrelevant ones. Besides, as users always have similar attention to the same events and most of the users' attentions imply the latent viewpoints of the events to some extent, we add user factor on the basis of traditional analysis of texts and images to mine the latent viewpoints (i.e., subevents). Furthermore, a multimedia subevent summarization process is designed to identify both representative texts and images, which are further aggregated to form a holistic visualized summary for the events.

In this paper, we focus on the event summarization process after social event detection: given the microblog dataset, we target at constructing event-related dataset and mining subevents as well as summarizing the subevents from both textual and visual aspects. Specifically, the proposed framework comprises three stages: (1) preprocessing, (2) subevents mining, and (3) event summarization.

The main contributions of this paper are as follows:

(1) We expand the event name to inform a small amount of related words to represent the events. We proposed a coarse-to-fine filtering method to distinguish the relevance and irrelevance, and then eliminate the irrelevance that reduces the dataset's effect.

(2) We analyzed how the user attributes such as background and microblogging attentions affect the quality of subevent mining. And we also use three modalities of data—user, text, and image to reinforce the co-clustering results. The different viewpoints of an event can be predicted by the users' attentions. We propose a new user-cluster based subevent determination approach, which simultaneously and respectively clusters the users, texts, and images into different clusters and find the subevents for a given event based on user clusters.

(3) We use text and image summarization method to assist users to gain a more visualized understanding of the events, and use comprehensive analysis of texts and images to offer users text description and illustration. We consider the content similarity, significance, and diversity three measurements to ensure the representativeness of textual summary. The content similarity ensures its relevance to event, significance ensures the popularity, and diversity ensures diversity. While for image summarization, we take visual similarity, significance and diversity into consideration to rank images.

The remainder of the paper is as follows. In Section 2, we describe existing methods for breaking events processing on Twitter and Weibo. In Section 3, we give an overview of the proposed system. In Section 4, we introduce preprocessing for noise and irrelevance elimination. In Section 5, we give the proposed user-based subevents discovery approach. In Section 6, we provide the detail event summarization method. Comparison experiments of our approaches and discussion are given in Section 7. Finally, in Section 8, we conclude and discuss the future work.

2. Related works

In recent years, there are many works performed on Twitter and Weibo. Some works aimed at breaking events detection [3–19] and event location estimation [20–22]. However, the amount of works that have taken images and user information afflicted by the microblogs into consideration is small. In this section, we mainly introduce some research on Weibo and Twitter in recent years.

2.1. Event detection

Recently detecting emerging topics on social or news streams has been a hot area both for industrial and academic communities. Zhou et al. [3] proposed a framework to detect composite social events over streams. They used location–time constrained topic (LTT) model to capture the content, time, and location of social messages, and represented each message as a probability distribution. Then they made use of the probability distribution to measure the similarity between messages. Nevertheless, the problem caused by texts, time and location of the tweets are not addressed in their model. Besides, tweet images are ignored in their work. Doman et al. [4] presented an event detection method based on "Twitter Enthusiasm Degrees (TED)" to generate a highlight video of a sports game. However, they ignored considering viewers' viewpoint in more depth by analyzing the criteria for including events in a highlight video. Cui et al. [6] utilized the hashtags as an indicator of events. They used three attributes of hashtags including instability, Twitter meme possibility, and

authorship entropy to discover breaking events. They should investigate more features for better measurement of the attributes, to discover more on the ambiguous hashtags. McMinn et al. [7] proposed a methodology for the creation of large-scale event detection corpora using state-of-the-art event detection approaches, and the Wikipedia current events portal to create a pool of events for further research and development. However, since events are given in prose, they cannot be compared automatically to results of event detection techniques. Ardon et al. [8] presented the comprehensive characterization of the diffusion of ideas on Twitter and performed a rigorous temporal and spatial analysis. Sakaki et al. [9] produced a probabilistic spatiotemporal model for the target event that can find the center and the trajectory of the event location. Popescu et al. [10] introduced some regression machine learning models to formalize the task of controversial event detection. Mathioudakis et al. [11] presented a system that performs trend detection over the Twitter stream. The system identified emerging topics (i.e. 'trends') on Twitter in real time and provided meaningful analytics that synthesize an accurate description of each topic. Aiello et al. [12] compared six trending topics detection methods, and proposed a topic detection method based on *n*-grams occurrence and topic ranking, which achieved the best performance. However, they could not detect the more interesting topics occurring within the event. Qu et al. [31] conducted content analysis of microblog messages, trend analysis of different topics, and an analysis of the information spreading process to investigate how Chinese netizens used microblogging in response to a major disaster: the 2010 Yushu Earthquake. Some other researchers focused on location estimation for breaking events. For example, Ozdikis et al. [20] applied an evidential reasoning technique in order to estimate the geographical locations of events in Twitter.

2.2. Event summarization

In recent years, there emerges many social media based applications, such as place-of-interest mining based on location information [32], personalized travel recommendation based on user interest and travel history [33], location estimation [21,22] and brand recommendation [34], video content summarization [35–38]. Many works have focused on event summarization based on the large amount of social media information: texts, images, user factors, location information. There have been some indeed efforts for providing vivid and attractive content for events.

Establishing complete event detection and visualization interface is one of the hottest topics in event summarization [14,39–41]. Gao et al. [14] presented a system called "GeSoDeck", which not only detected events, but also constructed a user interface to demonstrate both the detected event list for a geo-area and the images related to the event. Kuang et al. [39] proposed to carry out image and text visualization for events in microblogging services. Previous works have considered visualization contents to find meaningful information using spatiotemporal metadata, but they have lacked interaction and visualization to maintain the overall context. McMinn et al. [40] constructed an interface, which can automatically identify interesting events in real time, and gave a detailed overview as well as summarized information about events in a clean and easy-to-use interface. Shah et al. [41] presented a system that enables people to automatically generate a summary for a given event in real time by visualizing different social media such as Wikipedia and Flickr. However, they mainly relied on spatial-time analysis, and overlooked the correlations between different data types.

Some work just studied on document summarization [13,23,42]. Popescu et al. [13] proposed a method for automatically detecting events involving known entities from Twitter and both understanding the events and the audience reaction to them. However, they only found the audience opinions and main entities' actions without image visualization. Wu et al. [23] proposed an unsurprised method to summarize microblog by cascading two key-bigram extractors based on text rank and Latent Dirichlet Allocation (LDA). This method also ignored visual information on social media platform. Wang et al. [42] summarized events based on the minimum description length principle, which is achieved through learning a HMM from the event data.

While some works concentrated on text and image elaboration. For instance, Wang et al. [43] proposed a bilateral correspondence LDA model to address the problem of association modeling in multimedia microblog data that is to discover both text-to-image and image-to-text correspondence. However, this work is based on the images in Flickr, which cannot be well generalized to the ordinary Web images without social in-formation such as interest groups. Cai et al. [44] proposed a novel topic model which jointly used five Twitter features including text, image, location, timestamp and hashtag to mine breaking events, and used representative images selection to perform event visualization. However, this works neglected the fusion of text and image to offer a comprehensive and representative textual and visual summarization for events.

In addition, Schinas et al. [45] proposed a new visual event summarization method. They considered the textual, visual, social and time similarity between different images, and a multiple graph was generated to mine different topic under an event. However, the gap between the low-level features and the real semantics limits the model fidelity. Bian et al. [1,2] proposed a multimedia microblog summarization system called Cross-Media-LDA to automatically generate visualized summaries for trending topics. This is the most related work to ours, because it aimed to mine subevents in a given event. Unlike CMLDA or clustering between different data(usually texts and images), we introduce user feature into traditional co-clustering, which not only improve the diversity of event summarization, but also reinforce the similarity between texts and texts, images and images using similarity propagation. In this paper, we also propose a new framework called UTICC to mine subevents for a given event, and then use text and image summarization to select representative texts and images.

3. System overview

The object of our framework is to automatically generate a multimedia summarization (i.e., both textual and visual) from the Weibo dataset by mining subevents of the event. The flowchart of the proposed social media based event summarization framework is illustrated in Fig. 2. It consists of three main stages: (1) preprocessing, (2) subevent discovery, and (3) event summarization.

In the first stage, we propose a coarse-to-fine filtering method to reserve relevant data and eliminate noisy texts and images. At the same time, an event-related words generation method is proposed to mine related words for the events. We carry out coarse filtering by means of the initial query words of a known event, and then expand these words to construct event-related words set. Then we use event-related words to perform fine filtering for texts, and we also use the reinforced correlation between text and image to eliminate noisy photos.

In the second stage, a cross-media co-clustering model, termed User–Text–Image Co-clustering (UTICC), is proposed to jointly discover subevents from microblogs of multiple media types—user, text, and image. Based on inter-user interest similarity, the intrinsic correlations among these different data types are well explored and used for reinforcing the cross-media subevent discovery process. We also rank user clusters by their importance, and we find the most relevant text and image clusters to mining the latent subevents for each user cluster.

Finally, we propose an event summarization process to jointly identify representative texts and images for each subevent, which

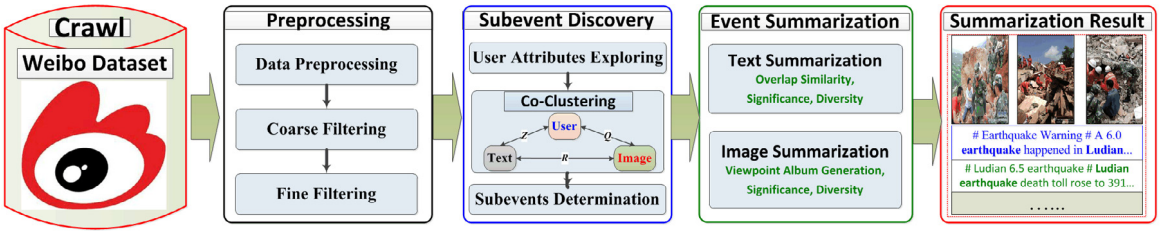

Fig. 2. Flowchart of our social media based event summarization system.

are further aggregated to form a holistic visualized summary. Specifically, we consider content similarity, significance and diversity to select representative texts, and we consider visual similarity, significance and diversity to choose representative images. In the end texts and images constitute the event summarization.

4. Preprocessing

We crawl a large collection of data from Weibo using its API.[1] The crawled information is displayed in Fig. 1. However, there exist much irrelevant information both in texts and images. What we need do is to reserve the relevance and remove the noise for a given event. In this section, we elaborate the details of preprocessing, including (1) data preprocessing, (2) coarse filtering, and (3) fine filtering.

4.1. Data preprocessing

In order to reduce the amount of noise before event summarization, we need to do some preprocessing in advance, which includes text preprocessing and image preprocessing.

For text preprocessing, we need to carry out word segmentation and keyword extraction. In this paper, we adopt existing natural language processing tool FudanNLP[2] to do this for each text. For example, for a text (i.e., a microblog) "I heard that a riot had just happened in the Kunming Railway Station" is segmented as "I heard that/a riot/ had just happened in the /Kunming/Railway Station"(the word segmentation) or "just, Railway Station, riot"(the keyword extraction). At the same time, we use stop-word filter to remove some word with high frequency but no special meanings, such as "I", "Ah", and "We" etc, as well as some words with low frequency [33].

For images, we firstly extract SIFT (Scale Invariant Feature Transform) feature. Each of the 128-dimension SIFT descriptors of an image is quantized to a visual vocabulary by hierarchical quantization [46]. Then each image is represented by a 10 000-dimensional bag of visual word histogram.

How to find the relevant multimedia information for a given event from its texts, images and other user generated information is one of the key steps in event summarization. So we carry out a coarse-to-fine filtering to remove noise.

4.2. Coarse filtering

The coarse filtering is mainly by means of event name and time stamp. Based on the preprocessed dataset, we utilize the event name as initial query words to carry out dataset filtering. For instance, we regard "Kunming", "Railway Station", "Chop", "Event" as query words for *2014 Kunming Attack*.[3] Because images in Weibo are not tagged, it is hard to determine whether the images are relevant to the event or not only by their visual feature. To solve this problem, we assume that if the text is related to the event, then its affiliated images are also related. We retrieve all texts posted in certain time spans and choose texts containing any one of the query words, together with their images, user information, and location to construct candidate dataset for a given event. Assume that a candidate event contains M_0 texts (with P_0 different words), N_0 images, together with L_0 users and the affiliated information.

4.3. Fine filtering

In candidate dataset, it may still contains some noise, because some users may upload irrelevant images when posting microblogs. Although the text contains the event name, it is not confirmed that the text is related to the event. So we further perform fine filtering to obtain relevant texts and images.

4.3.1. Fine filtering for texts

We propose the query word expansion by co-occurrence words generation to find event-related words, and then perform fine filtering for texts based on the expanded words.

4.3.1.1. Event-related words generation. In text pre-processing, we have divided each text into some participles. Suppose that the initial input query for an event consists of q words. We expand the q words to find event related words set by calculating the importance IM of each word to the query as follows:

$$IM_i = \frac{1}{q}\sum_{j=1}^{q} WS_{ij}, i = 1, \ldots, P_0 \tag{1}$$

where WS_{ij} is the word similarity between the ith word w_i and the jth word w_j, which is computed as follow:

$$WS_{ij} = exp\left\{-\frac{\max\left(\log n\left(w_i\right), \log n\left(w_j\right)\right) - \log n\left(w_i, w_j\right)}{P_0 - min(\log n\left(w_i\right), \log n\left(w_j\right))}\right\} \tag{2}$$

where $n\left(w_i\right)$ is the occurrence number of w_i in all texts, $n\left(w_i, w_j\right)$ is the co-occurrence number of w_i and w_j.P_0 is the total number of all words. The bigger the WS is, more similar the words are. We sort IM in descending order, and automatically choose the top P words with higher scores to construct the event-related words set, which is denoted by S.

$$P = \arg\max_i \{|IM_i - IM_{i+1}|\} \tag{3}$$

where P is the total number of event-related words, i is the word's index that has the maximum difference with its next word after sorting.

4.3.1.2. Fine filtering for texts. After getting the events-related words S which consists of P words, but how to use them to find relevant texts for describing the event? Different word has different contribution for finding relevant texts. Assume that there are n words in each text. The relevance of a text to the event

[1] http://open.weibo.com/wiki/API.

[2] https://code.google.com/p/fudannlp/downloads/list.

[3] http://weibo.com/p/100808c9af26add2e6a97a74680952d7abded2?feed_sort=hot&feed_filter=hot#Pl_Third_App__9.

(denoted by TR) can be measured by the average importance score of the n words, which is defined as follows:

$$TR = \frac{1}{n}\sum_{i=1}^{n} IM_i \tag{4}$$

The way of selecting the top M texts with higher relevance scores are similar to Eq. (3), and they construct event-related texts set.

4.3.2. Fine filtering for images

In order to find relevance images, we reinforce the correlation between text and image to bridge the semantic gap [47]. The basic idea is that, if the nearest neighbors (determined by the similarity of their word vectors) of each text are associated to an image, and then the text itself is more likely to be associated to the image. Likewise, if the nearest neighbors (determined by the similarity of their feature vectors) of each image are associated to a text, this image is more likely to be related to the text.

We calculate the bag-of-words for all M_0 texts. Then a text j can be represented as a P_o dimensional binary word vector $F_j = [F_{jk}]_{k=1}^{P_o}$, where $F_{jk} = 1$ represents the corresponding word k appears in this text, and $F_{jk} = 0$ represents it does not appear. Meanwhile, we can obtain the textual similarity TS of two texts through the cosine of their word vectors.

$$TS_{ij} = \langle F_i \cdot F_j \rangle / (|F_i| \cdot |F_j|) \tag{5}$$

where F_i denotes the word vector of the ith text, $|v|$ denotes the norm of the vector v, and "$\cdot$" denotes dot product. In this paper, the visual similarity VS_{ij} between the ith image and jth image is the cosine similarity of their bag of visual word histograms, which is computed similar to TS [46].

We firstly calculate the initial correlation matrix R' between text and image. Each element R'_{ij} in matrix R' represents whether the ith ($i = [1, M]$) text is related to the jth image ($j = [1, N_0]$). We set $R'_{ij} = 1$, if the jth image is associated with the ith text, otherwise $R'_{ij} = 0$. Then we reinforce the correlation between text and image [47], denoted by R'', as follows:

$$R''_{ij} = R'_{ij} + \frac{\sum_{k\in NN_t(t_i)} TS_{ik}R'_{kj}}{\sum_{k\in NN_t(t_i)} TS_{ik}} + \frac{\sum_{k\in NN_p(p_j)} VS_{jk}R'_{ik}}{\sum_{k\in NN_p(p_j)} VS_{jk}} \tag{6}$$

where $NN_t(t_i)$ is the k nearest neighbors of the ith text, $NN_p(p_j)$ is the k nearest neighbors of the jth image. TS_{ik} is the textual similarity between the jth image and its kth neighbor, and VS_{jk} is the visual similarity between the jth image and its kth neighbor. Then we get the normalized correlation matrix R_{ij} between the ith text and the jth image, which is also regarded as the joint probability matrix of text and image.

$$R_{ij} = \frac{R''_{ij}}{\sum_{i,j} R''_{ij}} \tag{7}$$

Then the relevance of each image IR can be computed as:

$$IR_i = TR_j \times R_{ij} \tag{8}$$

where TR_j is the relevance score of its corresponding text, R_{ij} is the association value between text j and image i. We use the similar method for words selection to select the top N images. Thus, the final event-related dataset contains M texts, N images, and L users in total.

5. User–text–image co-clustering

After constructing event-related dataset, we have obtained many relevant texts and images with their affiliated user information and so on. Now we turn to mine subevent by user–text–image co-clustering.

Co-clustering has recently received a lot of attention because it is a good method to simultaneously cluster heterogeneous correlated modalities [25,26,48–50]. We innovatively add user attributes on the basis of traditional co-clustering method, and perform UTICC to summarize the event by extending the cross-platform multimedia co-clustering framework [25].

On the one hand, since user, text and image are pair-wise dependent, we can use the relationship between user and text and the relationship between user and image to reinforce the clustering results of texts and images. On the other hand, based on the clustered users, we can find the most relevant text clusters and images clusters to construct subevents. Since the users in different clusters are dissimilar, the obtained subevents are diverse to each other, which makes the summary of the events diverse.

Our approach consists of the following three steps: (1) user attributes exploring, (2) user–text–image co-clustering, and (3) subevents determination.

5.1. User attributes exploring

User information plays an important role in social media analysis and recommendation [5,21,23,32–34,49–52]. Recommendation systems that take the user factor into account can be more personalized. Here we compute the user similarity based on his/her background information and microblogging information. And user similarity is further used to obtain the joint probability matrices between user and text, user and image.

5.1.1. Background similarity

We consider the background information of each user: gender, microblogging activity information (includes followers' number, attentions number, and microblogs number), and location information, in our event summarization. As we know, users with similar background information are more likely to have similar interest and concern. We compute the background similarity as follows:

$$B_{ij} = b_g + b_a + b_l \tag{9}$$

where b_g represents the gender similarity, we set $b_g = 1$ if two users have same gender, otherwise $b_g = 0$. b_a is the microblogging activity similarity, which is determined by the cosine of the microblogging activity information of the two users. b_l is the location similarity. We set b_l by considering the province level and city level as follows:

$$b_l = \begin{cases} 1, & if\ in\ the\ same\ city \\ 0, & if\ not\ in\ the\ same\ provice \\ 0.5, & otherwise \end{cases} \tag{10}$$

5.1.2. Microblogging attention similarity

Here, we mine users interest mainly from the texts and images uploaded by the users. We compute the average textual similarity and visual similarity to represent users' microblogging attention similarity A_{ij} as follows:

$$A_{ij} = \frac{w_1}{nt_i \times nt_j}\sum_{m=1}^{nt_i}\sum_{n=1}^{nt_j} TS_{mn} + \frac{w_2}{np_i \times np_j}\sum_{m=1}^{np_i}\sum_{n=1}^{np_j} VS_{mn} \tag{11}$$

where nt_i and np_i respectively represent the total text and image number of user i. TS_{mn} is the textual similarity of text m of user i and text n of user j, and correspondingly VS_{mn} is the visual similarity of image m and image n. w_1 and w_2 are respectively the weights of textual similarity and visual similarity. Some users have uploaded images, while others have not. So w_1 and w_2 are depending on the specific circumstances. When a user both use text and image to describe events, we set $w_1 = w_2 = 0.5$. When user only has texts, we set $w_1 = 1, w_2 = 0$.

Finally, we represent the similarity of two users i and j denoted by US_{ij} by taking background similarity and user attention similarity into account:

$$US_{ij} = \delta_1 B_{ij} + (1 - \delta_1) A_{ij} \tag{12}$$

where δ_1 is the weight of background similarity, and $\delta_1 \in [0,1]$.

5.2. Co-clustering

As user, text and image these three data types are pair-wise dependent, we use co-clustering method to simultaneously clusters them. Here, we introduce user factor on the basis of traditional text and image co-clustering. Similar users are more likely to hold the same viewpoints of the events, thus for the clusters obtained by UTICC, data in the same clusters will be tighter and more similar, and data in different clusters will be more diverse.

From above we get the event-related dataset which includes M texts, N images and L users. The notations are listed in Table 1. In this section we simultaneously group the user set $\mathrm{U} = \{u_i\}_{i=1}^{n_U}$ into K clusters $\hat{U} = \{\hat{u}_k\}_{k=1}^{K}$, group text set $\mathrm{T} = \{t_i\}_{i=1}^{n_T}$ into G clusters $\hat{T} = \{\hat{t}_g\}_{g=1}^{G}$, and group image set $\mathrm{P} = \{p_i\}_{i=1}^{n_P}$ into H clusters $\hat{P} = \{\hat{p}_h\}_{h=1}^{H}$ by UTICC.

5.2.1. Preliminary

In order to carry out UTICC, we are required to respectively determine the joint probability matrix of text and image (denoted by R, as shown in Eq. (5)), user and text (denoted by Z), image and user (denoted by Q) respectively. The calculation of Z and Q is similar to R's.

Let respectively set Z', Z'', and Z as the initial correlation matrix, the reinforced matrix, and the joint probability matrix between user and text. We firstly calculate Z', and each Z'_{ij} in matrix Z' represents whether the ith ($i = [1, L]$) user is related to the jth text ($j = [1, M]$) by assigning 1 to them if the ith user has posted the jth text, otherwise assigning 0. Then we compute the reinforced correlation matrix Z'' as follows:

$$Z''_{ij} = Z'_{ij} + \frac{\sum_{k \in NN_u(u_i)} US_{ik} Z'_{kj}}{\sum_{k \in NN_u(u_i)} US_{ik}} + \frac{\sum_{k \in NN_t(t_j)} TS_{jk} Z'_{ik}}{\sum_{k \in NN_t(t_j)} TS_{jk}} \tag{13}$$

where $NN_u(u_i)$ and $NN_t(t_i)$ represents the k nearest neighbors of u_i and t_i. US_{ik}(or TS_{ik}) represents the user similarity (or textual similarity) between the ith user (or text) and its kth neighbor. The value of k is Manually specified.

Then we get the joint probability matrix Z_{ij} between the ith user and the jth text as follows:

$$Z_{ij} = \frac{Z''_{ij}}{\sum_{i,j} Z''_{ij}} \tag{14}$$

Let respectively set Q', Q'', and Q as the initial correlation matrix, the reinforced matrix, and the joint probability matrix between image and user. We firstly calculate Q', and each Q'_{ij} in matrix Q' represents whether the ith ($i \in [1, N]$) image is related to the jth user by assigning 1 to them if the jth user has posted the ith image, otherwise assigning 0. And then we compute the reinforced correlation matrix Q'' as follows:

$$Q''_{ij} = Q'_{ij} + \frac{\sum_{k \in NN_p(p_i)} VS_{ik} Q'_{kj}}{\sum_{k \in NN_p(p_i)} VS_{ik}} + \frac{\sum_{k \in NN_u(u_j)} US_{jk} Q'_{ik}}{\sum_{k \in NN_u(u_j)} US_{jk}} \tag{15}$$

where $NN_p(p_i)$ represents the k nearest neighbors of p_i. VS_{ik} represents the visual similarity between the ith image and its kth neighbor. The value of k is Manually specified.

The joint probability matrix Q is computed as follows:

$$Q_{ij} = \frac{Q''_{ij}}{\sum_{i,j} Q''_{ij}} \tag{16}$$

5.2.2. UTICC

Based on the joint probability matrix of text and image R, user and text Z, and image and user Q, we propose a user–text–image co-clustering approach by finding the optimal user, text and image mapping $(\rho^*, \sigma^*, \varphi^*)$ that minimizes the linear combination of the Bregman information of Z, R and Q as follows:

$$\begin{aligned}(\rho^*, \sigma^*, \varphi^*) &= \underset{\rho,\sigma,\varphi}{\arg\min}\{\alpha I_{\varnothing}(Z) + \beta I_{\varnothing}(R) + \gamma I_{\varnothing}(Q)\} \\ &= \underset{\rho,\sigma,\varphi}{\arg\min}\Big\{\alpha \sum_{u:\rho(u)=\hat{u}} \sum_{t:\sigma(t)=\hat{t}} p_{ut} d_{\varnothing}\left(z_{ut}, \tilde{z}_{ut}\right) \\ &\quad + \beta \sum_{t:\sigma(t)=\hat{t}} \sum_{p:\varphi(p)=\hat{p}} p_{tp} d_{\varnothing}\left(r_{tp}, \tilde{r}_{tp}\right) \\ &\quad + \gamma \sum_{p:\varphi(p)=\hat{p}} \sum_{u:\rho(u)=\hat{u}} p_{pu} d_{\varnothing}\left(q_{pu}, \tilde{q}_{pu}\right)\Big\}\end{aligned} \tag{17}$$

where $\tilde{Z}, \tilde{R}, \tilde{Q}$ are the approximation matrix of Z, R and Q, which are obtained by the generalized maximum entropy approach [26, 48]. $I_{\varnothing}(\cdot)$ is the Bregman information, and $d_{\varnothing}(\cdot)$ is the Bregman divergence [26,48]. And when optimizing the object function (17), we fix σ, φ to choose the best ρ. Likewise, we fix ρ, φ to choose the best σ and fix ρ, σ to choose the best φ [25]. The pseudocode for UTICC is shown in **Algorithm 1**.

After UTICC, we respectively obtain user cluster s $\hat{U}$, text.

We also obtain the final approximation matrix $\tilde{Z}, \tilde{R}, \tilde{Q}$ from which we can determine subevents.

Algorithm1: User–Text–Image Co-clustering Algorithm (UTICC)

Input	User, text and image set U, T, P the joint probability matrix Z, R, Q joint probability p_{ut}, p_{tp}, p_{pu} parameter α, β, γ, Bregman divergence $d_{\varnothing}$
Output	Clusters users, texts and images $\hat{U}, \hat{T}, \hat{P}$ approximation matrix $\tilde{Z}, \tilde{R}, \tilde{Q}$ co-clustering ρ, σ, φ

Initialization: a random co-clustering ρ, σ, φ based on U, T, P
Repeat:
Update $\hat{U}, \hat{T}, \hat{P}$ and compute $\tilde{Z}, \tilde{R}, \tilde{Q}$
For each user, find its new cluster as follows, and update $\hat{U}$.

$$\hat{u}^* = \underset{\hat{u}}{\operatorname{argmin}} \alpha \sum_{\hat{t}} \sum_{\sigma(t)=\hat{t}} p_{ut} d_{\varnothing}\left(z_{ut}, \tilde{z}_{ut}\left(\hat{u}\right)\right) + \beta \sum_{\hat{u}} \sum_{\rho(u)=\hat{u}} p_{pu} d_{\varnothing}(q_{pu}, \tilde{q}_{pu}(\hat{p}))$$

For each text, find its new cluster as follows, and update $\hat{T}$.

$$\hat{t}^* = \underset{\hat{t}}{\operatorname{argmin}} \alpha \sum_{\hat{u}} \sum_{\rho(u)=\hat{u}} p_{ut} d_{\varnothing}\left(z_{ut}, \tilde{z}_{ut}\left(\hat{u}\right)\right) + \gamma \sum_{\hat{p}} \sum_{\varphi(p)=\hat{p}} p_{tp} d_{\varnothing}\left(r_{tp}, \tilde{r}_{tp}\left(\hat{p}\right)\right)$$

For each image, find its new cluster as follows, and update $\hat{P}$.

$$\hat{p}^* = \underset{\hat{p}}{\operatorname{argmin}} \beta \sum_{\hat{t}} \sum_{\sigma(t)=\hat{t}} p_{tp} d_{\varnothing}\left(r_{tp}, \tilde{r}_{tp}\left(\hat{p}\right)\right) + \gamma \sum_{\hat{u}} \sum_{\rho(u)=\hat{u}} p_{pu} d_{\varnothing}(q_{pu}, \tilde{q}_{pu}(\hat{p}))$$

Until convergence
Return $\tilde{Z}, \tilde{R}, \tilde{Q}, \rho, \sigma, \varphi, \hat{U}, \hat{T}, \hat{P}$

Table 1
List of key notations.

Notation	Description
U, T, P	User set, text set, and image set, $U = \{u_l\}_{l=1}^{L}$, $T = \{t_m\}_{m=1}^{M}$, $P = \{p_n\}_{n=1}^{N}$
L, M, N	The number of users, texts, and images
$\hat{U}, \hat{T}, \hat{P}$	Clustered user, text, and image,$\hat{U} = \{\hat{u}_k\}_{k=1}^{K}$, $\hat{T} = \hat{T} = \{\hat{t}_g\}_{g=1}^{G}$, $\hat{P} = \{\hat{p}_h\}_{h=1}^{H}$
K, G, H	The number of user cluster, text cluster, and image cluster
Z', R', Q'	Initial matrix between user and text, text and image, image and user
Z'', R'', Q''	Reinforced probability matrix between user and text, text and image, image and user
Z, R, Q	Joint probability matrix between user and text, text and image, image and user
$\tilde{Z}, \tilde{R}, \tilde{Q}$	Approximation matrix of Z, R, Q, $\tilde{Z} = [\tilde{z}_{ut}]$, $\tilde{R} = [\tilde{r}_{tp}]$, $\tilde{Q} = [\tilde{q}_{pu}]$
p_{ut}, p_{tp}, p_{pu}	Joint probability distribution between user and text, text and image, image and user
ρ, σ, φ	The mapping of user, text and image, i.e.,$\rho: U \to \hat{U}$, $\sigma: T \to \hat{T}$ and $\varphi: P \to \hat{P}$

5.2.3. Post processing for UTICC

There may exist small clusters that only contains few data, so we carry out some post-processing for the clusters $\hat{U}, \hat{T}, \hat{P}$. We first prune the clusters whose number is smaller than a certain threshold, which is empirically set as $0.05 \times L$ (for $\hat{U}$), $0.05 \times M$ (for $\hat{T}$) and $0.05 \times N$ (for $\hat{P}$). In addition, some "closest" pair of clusters, between which the user similarity (for $\hat{U}$), textual similarity (for $\hat{T}$) or visual similarity (for $\hat{P}$) is bigger than 0.5, will be merged together into a new cluster. After these processes, the clusters reserved are used to determine subevents in the next section.

5.3. Subevents determination

The most similar users have been gathered together, and these users' attentions are more representative in a given event. Hence, we determine subevents based on the user clusters.

The approximation matrix $\tilde{Z}$, and $\tilde{Q}$ are used to measure the correlation between $\hat{U}$ and $\hat{T}$, $\hat{U}$ and $\hat{P}$. We set the minimum number of the clusters of user, text and image as the number of final subevents L, i.e., $E = \min(K, G, H)$. Our subevent determination approach consists of the following two steps. Firstly, we rank the user clusters by their importance. And the importance of each user cluster is related to the number of users in it.

Then, for each user cluster $\hat{u}$, we determine a subevent $[\hat{t}, \hat{p}]$ by choosing the most relevant text cluster $\hat{t}$, and the most relevant image cluster $\hat{p}$ as follows:

$$\begin{aligned} \hat{t}^* &= \arg\max_{\hat{t}} \{\tilde{z}_{ut}\}, \hat{t} \in \hat{T} - \Omega_T \\ \hat{p}^* &= \arg\max_{\hat{p}} \{\tilde{q}_{pu}\}, \hat{p} \in \hat{P} - \Omega_P \end{aligned} \tag{18}$$

where $\hat{T}$ and $\hat{P}$ are the whole text clusters and image clusters. Ω_T and Ω_P are the clusters set that have been chosen when determining subevents. We initially set $\Omega_T = \varnothing$, $\Omega_P = \varnothing$, and we set $\Omega_T = \Omega_T + \hat{t}^*$, $\Omega_P = \Omega_P + \hat{p}^*$ after constructing a subevent.

We continue the subevent generation by choosing new text cluster $\hat{t}$ and picture cluster $\hat{p}$ from the remaining user clusters until there is no user clusters. And finally we construct L subevents, and there are many texts and images in each subevent.

6. Event summarization

As for each subevent, we have selected the most relevant text clusters $\hat{t}$ and image clusters $\hat{p}$. However, there are many texts in $\hat{t}$ and images in $\hat{p}$. In this section, we logically carry out representative texts and images selection for each subevent. Our event summarization approach consists of the following two aspects: text summarization, and image summarization.

6.1. Text summarization

For each text cluster $\hat{t}$, we rank the texts in it according to the content similarity, diversity and significance. Specifically, we consider content similarity to ensure the relevance to the event, diversity to offer diverse summary, and significance to choose more popular and significant texts.

6.1.1. Content similarity

We measure content similarity of texts by a recall-liked score [23], which counts the overlap words between the text and the event-related words S. Formally, we computes CS_j score of a text t_j as follows:

$$CS_j = \frac{|\{w_i | w_i \in t_j \& w_i \in S\}|}{\max\left(len\left(t_j\right), AvgLen\right)} \tag{19}$$

where w_i is the word in text t_j, $len\left(t_j\right)$ is its length, and $AvgLen$ is the average length of all the texts in event-related dataset. $|\{w_i | w_i \in t_j \& w_i \in S\}|$ is the number of co-occurrence words between text t_j and S.

We sort the texts by their content similarity scores, and the top ranking texts are more relevant and can well represent the events. However, similar texts may get the similar scores, so, we should take the diversity into consideration.

6.1.2. Diversity

For a text t_j in a subevent, we measure its diversity score DS_j by the dissimilarity among it and all the other texts in the same subevent as follows:

$$DS_j = \frac{1}{n_t} \sum_{i=1}^{n_t} \left(1 - TS_{ij}\right) \tag{20}$$

where n_t is the text number in $\hat{t}$, and TS_{ij} is the text similarity between the ith text and text t_j as shown in (2). Texts with higher DS are more diverse than others.

6.1.3. Significance

In addition, each microblog can be commented, reposted and praised (in Chinese: 点赞) by users. And these social behaviors imply the popularity of different microblogs. Let com_j, rep_j, and pra_j respectively denote the corresponding number of comment, repost and praise of a text t_j. We use a smooth function over the number to measure the significance of a text as follows:

$$Sig_j = log(com_j + rep_j + pra_j) \tag{21}$$

6.1.4. Representative text selection

We aim to find the texts with higher content similarity, significance, diversity score, so we use the linear combination of these three scores.

$$score_j = \tau_1 CS_j + \tau_2 DS_j + \tau_3 Sig_j \tag{22}$$

Table 2
Some related words and texts of ludian earthquake.

Ludian	Earthquake	Ludian country	Seismological Bureau
Great earthquake	Earthquake area	Ludian people	Care Ludian
An earthquake of magnitude 6.5 occurred in ludian, yunan province			
Very sad about yunnan earthquake. Pity for those people. Why should there be an earthquake			
Anyway, come on, yunnan! Come on ludian!			

where τ_1, τ_2 and τ_3 are trade-off parameters, each of them are in the range [0,1] and $\tau_1 + \tau_2 + \tau_3 = 1$. And each score OS_j, DS_j, Sig_j used here has been normalized in advance. We sort score in descending order, and texts with high score are chosen as representative text for each subevent. Here we simply set $\tau_1 = 0.6$, $\tau_2 = \tau_3 = 0.2$. More discussions are given in the subsequent experiments. In each subevent, we use the maximum difference to determine the final representative text number.

6.2. Image summarization

In this section, we aim to accomplish image summarization for the event. Note that, for each image cluster $\hat{p}$, if there are several images visually very similar and appearing many times, these images are more likely to be representative. In this paper, we first use visual grouping to group visual similar images. Then we use significance and diversity re-ranking to select the final representative images to summarize each event.

6.2.1. Visual grouping

We use graph growth based viewpoint album generation algorithm [46] to group visually similar images into different albums. Images in the same albums are visually alike. However, some albums only contain two or three images and they may have nothing to do with events, so we need to weed them out and only keep albums whose number is bigger than a certain number. In each album, we choose an image that is the most similar one to others in the album. Thus, we only reserve a few numbers of images in each subevent as the representative images.

6.2.2. Re-ranking

After visual analysis, we also compute the significance and diversity score for each image to select representative images for the event.

The significance of text can also be regarded as that of its affiliated images. While diversity is similar to (20), we just use visual similarity to replace textual similarity. We compute the linear combination of significance and diversity, and apply the similar method like representative texts selection to choose images.

6.3. An example

Here we make Ludian Earthquake as an example to well explain the proposed approach. For text preprocessing, we adopt FudanNLP and a text "I heard that an earthquake had just happened in Ludian" is segmented as "I heard that/an earthquake/had just happened in/Ludian"(the word segmentation) and "just, Ludian, earthquake"(the keyword extraction).

Then we carry out event-related words generation and fine filtering for texts. Table 2 shows some examples of event related words and texts.

Based on the event related words, we calculate similar texts. For example, "An earthquake of magnitude 6.5 occurred in ludian, yunan province" is very similar to "There was an earthquake of magnitude 6.5 in ludian, yunan province". Based on the initial correlation matrix between text and image and similarity matrix of texts, we reinforced the correlation between one text and other images. Table 3 is an example. Then we carry out event related images.

Table 3
An example of initial and reinforced correlation of one text.

The Red Cross society of China has received more than 38 million yuan in donations within 24 hours of the ludian earthquake in yunnan province, said zhao baige.	
initial	
reinforced	

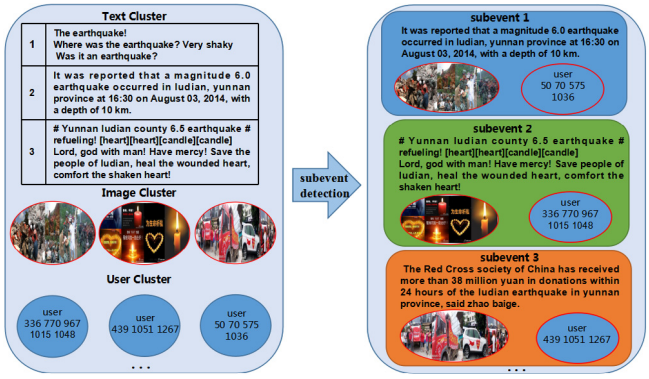

Fig. 3. Some examples of UTICC and subevent detection.

Table 3 is an example. Then we carry out event related images.

We calculate the user similarity based on his/her background information and microblogging information. We calculate the initial correlation matrix between user and text, user and image, then reinforce it. The approach is the same as 4.3.2). Then we carry out user–text–image co-clustering and subevents determination in Fig. 3. After text summarization and image summarization, the final result is shown in Fig. 10(b).

7. Experiments

In this section, we present the experiment results of the proposed framework of social media based event summarization, e.g., text summarization performance, image summarization performance and event-related words generation. In order to verify the effectiveness of our method, we perform a series of contrast experiments. The number of text and images we choose are respectively 50 and 10. The six methods are as follows:

Random: randomly select texts and images from the filtered set.

Kmeans: applies Kmeans for the filtered set, and picks up the nearest text and image in terms of distance to the centers. The mean value is 50 in text summarization and 10 in image summarization.

Sig-based: ranks the texts in the filtered set by their significance score, and picks up the more significant texts as text summary.

TRLDA (Text Rank LDA): Firstly, we use TextRank to extract the key double-conjoined words, then select representative texts and images by LDA classification [48]. The number of iterations is 300, the random seed is 3, beta = 0.01, alpha = 50(or 10)/topic number.

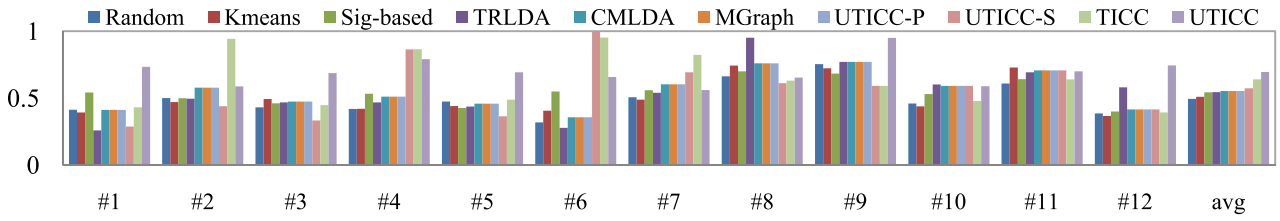

Fig. 4. Performance comparison for text summarization results using different methods. The *x* axis is event id, *y* axis is F-score of ROUGE-1.

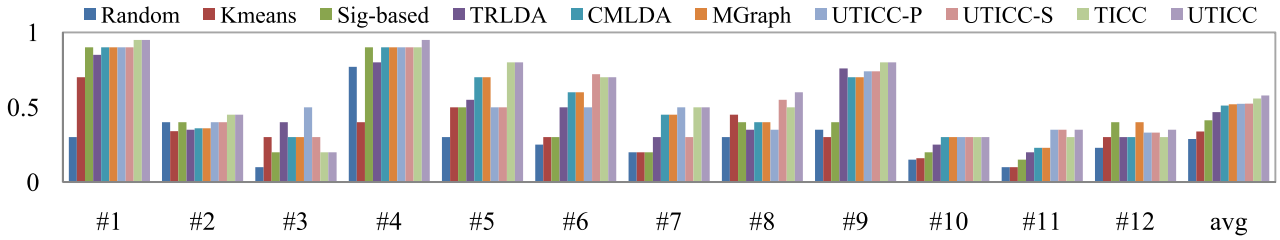

Fig. 5. Performance comparison for image summarization results using different methods. The *x* axis is event id, *y* axis is P@10.

CMLDA (Cross-Media-LDA): generates visualized summaries for trending topics according to [1]. The parameter setting is the same as TRLDA.

MGraph (topic Modeling and Graph-based ranking): use a topic modeling technique to capture the relevance of messages, and uses graph-based algorithm to produce a diverse image ranking [45].

UTICC: our proposed method in this paper.

UTICC-P: UTICC without preprocessing, i.e., we directly discover subevents in a given event, and perform multimedia summarization for it.

UTICC-S: UTICC without subevent discovery, i.e., we only perform preprocessing for dataset, and offer multimedia summarization for the events.

TICC (Text–Image Co-Clustering): uses TICC to discover the events, the most relevant text clusters and image clusters are merged to a subevent. The preprocessing and event summarization is the same to UTICC.

7.1. Dataset and experiment settings

In our preliminary data collection step, we crawled data from following seven cities in China, which are **Beijing**, **Shanghai**, **Guangzhou**, **Xi'an**, **Kunming**, **Urumqi**, and **Sanya**. Each city has its official geographical location information which can be consulted from Wikipedia. Based on the geographical location information of microblogs, we crawl all of the microblogs if their locations are within one kilometer to the seven cities. And the data range from March 2014 to April, 2015. The microblog information includes text, images, user id, and picture id, latitude, longitude, etc.

Another condition should be noted is that all the microblog data we crawled contain location information, i.e., microblogs without location information is not collected. In Table 4, we list the numbers of texts, users and images in our Weibo dataset. For instance, in Beijing, we have crawled 2776,623 texts, 4432,514 images, which are uploaded by 703,264 users, i.e., each user has posted 4 texts and 6 images on average. And about 38.71% of the texts have affiliated images.

In Table 5, some detailed information about the 12 events is given. Taken #1Kunming Station Massacre as an example, it occurred in the Kunming Railway Station on Mar 1, 2014.

When computing user similarity, we set $\delta_1 = 0.4$. And when performing user–text–image co-clustering, we empirically set $\alpha = 0.2$, $\beta = 0.1$, $\gamma = 0.7$, and $K = G = H = 20$ initially. The final K, G and H are unstable, which vary according to the event dataset.

Table 4
The numbers of texts, users and images in our dataset.

City name	Text No.	User No.	Image No.
Beijing	2776,623	703,264	4432,514
Shanghai	2282,225	553,170	3737,307
Guangzhou	2000,362	515,547	4443,056
Xi'an	880,134	214,041	1253,115
Kunming	540,590	144,533	908,611
Urumchi	427,711	88,529	676,272
Sanya	198,159	58,873	420,705
Total	9105,804	2277,957	15,871,580

Table 5
Some detailed information about the events.

EventID	Name	Starting Time	Location
#1	Kunming Station Massacre	Mar 1, 2014	Kunming Station
#2	MH370 Disappearance	August 5, 2014	Unknown
#3	Shanghai Stampede	Dec 31, 2014	Shanghai Bund
#4	Urumqi Attack	May 22, 2014	Urumqi
#5	Ludian Earthquake	Aug 3, 2014	Ludian, Yunnan
#6	Ice Bucket Challenge	August 25, 2014	Unknown
#7	Death of Siangtan Pregnant	August 10, 2014	Siangtan, Hunan
#8	Harbin Warehouse Fire	Jan 2, 2015	Harbin
#9	Wilson Typhoon	Jul 17, 2014	Hainan, Guilin, Guangzhou
#10	Zhaoyuan McDonald's Killing	May 28, 2014	Zhaoyuan, Shandong
#11	Foshing Airport Distress Landing	Jul 23, 2014	Foshing, Taiwan
#12	Explosion in Kaohsiung	Aug 1, 2014	Kaohsiung, Taiwan

7.2. Text summarization performance

In text evaluations, for fairness of evaluation, we manually choose 50 relevance texts from the event dataset, which construct the Ground Truth [1,2,23]. As this data is Chinese, we first use FudanNLP to perform word segmentation, then use them to compare the text summaries obtained by different methods. As we all know ROGUE is a well-known method for retrieval evaluation, we take

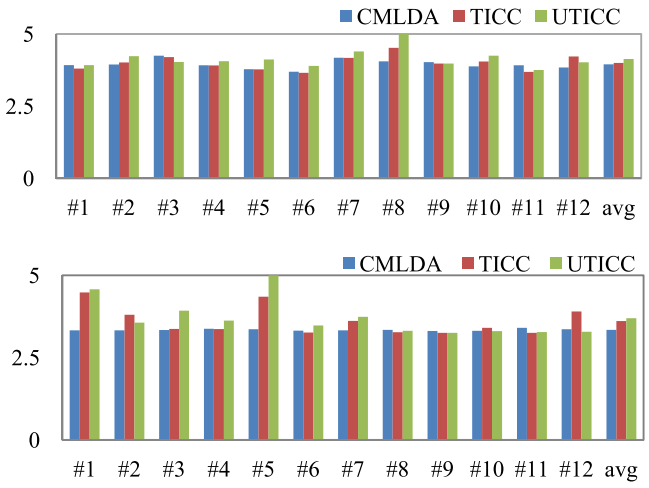

Fig. 6. Performance of intra-subevent relevance and inter-subevent diversity using CMLDA, TICC and UTICC. The figure above is intra-subevent relevance score, and the figure below is inter-subevent diversity score. The *x*-axis is event id, and *y*-axis is volunteers' score.

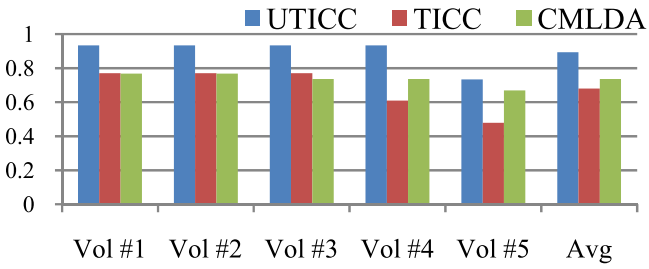

Fig. 7. The average Kappa value of 5 volunteers. The *x* axis is volunteer id, *y* axis is kappa value.

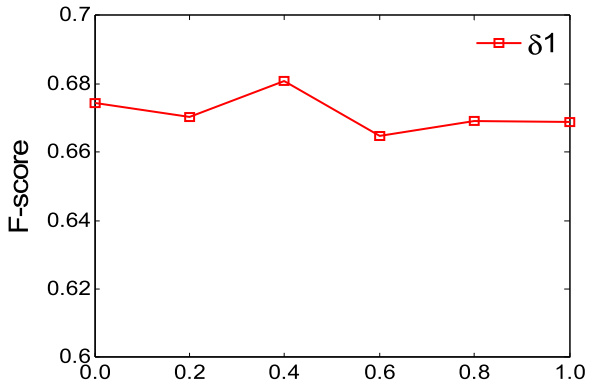

Fig. 8. F-score Performance of parameter δ_1 used for user similarity computation.

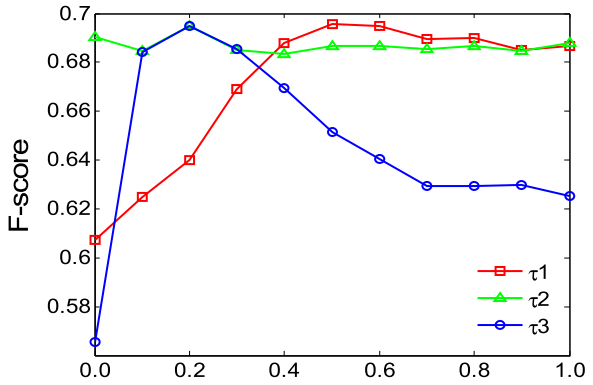

Fig. 9. Precision Performance of parameter τ_1, τ_2 and τ_3.

ROUGE-1 as an example [53]. Let GT be the Ground Truth, and AS be the automatic text summary obtained by different summarization methods. At first, Recall and Precision are computed as follows:

$$recall = \frac{\sum_{t \in GT} \sum_{w \in t} match\,(w)}{\sum_{t \in GT} \sum_{w \in t} count(w)} \tag{23}$$

$$precision = \frac{\sum_{t \in AS} \sum_{w \in t} match\,(w)}{\sum_{t \in AS} \sum_{w \in t} count(w)} \tag{24}$$

$$F - score = \frac{2 \times recall \times precision}{recall \times precision} \tag{25}$$

where $match(w)$ is the match numbers of words in GT and AS, $count(w)$ is the total word number in each text. The bigger the F-score is, the better the text summarization performance is.

The overall comparison of proposes TICC and UTICC with other existing approaches are shown in Fig. 4. In the last column, we also give the average performance of different approaches. We can see from the results, the proposed UTICC outperforms other method in majority of the events. The outstanding performance of the UTICC benefits from the following aspects.

First of all, UTICC explores the joint correlations between user, text and image. The impact of the multiple modalities can be illustrated by the results of TICC, which differs form UTICC only with the lack of user analysis. The comparing results show that the introduction of user attributes has improved the performance. In addition, by comparing the results of UTICC-P (without the preprocessing, i.e., noise elimination) and UTICC, it shows the contribution of preprocessing. UTICC works better than UTICC-S (without mining subevents), which illustrates the importance of subevent discovery. Although CMLDA considered the subevent discovery, it only used the joint correlation between textual and visual aspects of microblogs and ignored users' similar interests and common attentions, which is the largest difference to UTICC.

7.3. Image summarization performance

In image evaluations, we apply the proposed method to generate a representative summary for each of the 12 events. In particular, we rank the images and select the top N as summary. We evaluate the performance of UTICC, Sig-based, TRLDA and CMLDA in a similar manner as [54] by calculating the P@N, Mean Reciprocal Rank (MRR) and Average Visual Similarity (AVS@N).

P@N: the averaged percentage of images in top N that are relevant to the corresponding event. We calculate percentage when $N = 10$. Bigger P@N means better image summarization performance.

Mean Reciprocal Rank (MRR): This is computed as 1/r, where r is the average rank of the first relevant image.

Average Visual Similarity (AVS@N): This measures the average visual similarity among all pairs of images in top N. Lower AVS values implys higher diversity in terms of visual content.

The average performances of the 12 events among different image summarization results are reported in Table 6. We can see that UTICC gains the best performance on P@10 and MRR. The comparison of UTICC-P, UTICC-S, and TICC fully illustrates the advantages of preprocessing, subevent summary, and the application of user attributes. It should be noted that, for the AVS@10 comparison, Random gains a better performance than other approaches because the performance of Random is not stable. All in all, UTICC

(a)

Fig. 10. Event summarization on (a) event #1 Kunming Station Massacre, (b) event #5 Ludian Earthquake, (c) event #4 Urumqi Attack, (d) event #3 Shanghai Stampede. Images and texts in the same dotted box are corresponding to one subevent.

Table 6
Average performances of the 12 events among different image summarization methods.

Method	P@10	MRR	AVS@10
Random	0.2875	0.2894	**0.2325**
Kmeans	0.3375	0.3360	0.4250
Sig-based	0.4125	0.4033	0.3964
TRLDA	0.4675	0.7217	0.3198
CMLDA	0.5117	0.7778	0.3864
MGraph	0.5200	0.8608	0.3278
UTICC-P	0.5225	0.8750	0.3283
UTICC-S	0.5242	0.9167	0.3234
TICC	0.5583	0.9533	0.3328
UTICC	**0.5792**	**1.0000**	**0.3045**

works the best when considering the precision and the MRR, as well as the text summarization performance.

In order to obviously illustrate the improvements, we show the P@10 of image summarization in Fig. 5. Apart from event #3 Shanghai Stampede, our method works the best. This is because Shanghai Stampede simultaneity occurred with New Year Eve, these two events are all very popular. And for event #10–#12, the performances of all ten different methods are not good, this is because the image number is very small. And UTICC works better than the other three methods UTICC-P, UTICC-S, TICC, not only because of the introduction of user factor, but also the subevents discovery and the preprocessing for noise elimination.

(b)

Fig. 10. (*continued*).

7.4. Capability of user factor

In order to illustrate the importance of user factor, we perform some discussion experiments. The only difference between TICC to UTICC is that we only use text and image for co-clustering in TICC, and subevent determining and summarization method are the same. We invite 5 volunteers to evaluate the subevents summarization result in terms of two aspects: (1) intra-subevent relevance, (2) inter-subevent diversity. All of the volunteers are Chinese, and they are familiar with the events. Evaluations are categorized into five levels, i.e., {1, 2, 3, 4, 5} indicating "very bad", "bad", "average", "good", "very good". Finally we compute the average score of these 5 volunteers of intra-subevent relevance and inter-subevent diversity. The performance is shown in Fig. 6.

We can see in Fig. 6, UTICC obtains the best performance both on intra-subevent relevance and inter-subevent diversity. When using TICC, the performance is a little worse. This demonstrates that the introduction of user factor makes difference. In addition, our method also works better than CMLDA. This is because CMLDA only considers text and image in subevent mining, but ignores the user factor and the correlations between user and the other two types of data.

In order to measure the interrater reliability among 5 volunteers, we use SPSS to calculate the Kappa value between each two volunteers for 12 events, the result is shown in Fig. 7. As for UTICC, 4 of 5 volunteers' Kappa value and the average of all 5 volunteers are bigger than 0.8, which means almost perfect.

7.5. Discussions

In this section, we perform some discussions about parameters used in our method, e.g., user similarity parameter discussion, text summarization parameter discussion.

(c)

Fig. 10. (*continued*).

7.5.1. User similarity parameter discussion

When computing user similarity, we both consider background similarity B and attention similarity A. Meanwhile, we make a linear combination of them. So here we will discuss the effects of the weight δ_1 of B, while that of interest similarity is equal to (1-δ_1). We range δ_1 from 0 to 1 with the step of 0.2, and calculate the text performance when using different combination of B and A in similarity computation of UTICC. The performance comparison is shown in Fig. 8.

As we can see in Fig. 8 Text summary get the best performance when $\delta_1 = 0.4$. Starting from $\delta_1 = 0$ (i.e., we only use microblogging attention similarity), text summarization becomes better with the increase of δ_1. After $\delta_1 = 0.4$, the performance becomes a little worse. And when $\delta_1 = 1$ (i.e., we only use background similarity), the performance is the worst. So in our framework of event summarization, we choose 0.4 B + 0.6 A to compute user similarity by both considering background similarity and attention similarity.

In order to prove that the improvement is statistically significant, we carry out hypothesis test. We propose the hypothesis:

$$H_0: u_1 > u_2, H_1: u_1 \leq u_2$$

where u_1 is the mean value of F-score when $\delta_1 = 0.4$ u_2 is the mean value of F-score when $\delta_1 = 0$. Computation result shows that the hypothesis is accepted when significance level is 0.05.

7.5.2. Text summarization parameter discussion

The text selection score is a weighted linear combination of the three criteria—content similarity, significance and diversity. In this part, we examine the effects of the corresponding weighting parameters τ_1, τ_2 and τ_3. In order to achieve the optimal parameter setting, we use the parameter tuning method in [6], and the optimal parameters are selected as 0.6, 0.2, 0.2 for the events.

In order to prove that above results are the optimized combination, we further fix two of the values as the achieved value, and vary the third one. According to the results shown in Fig. 9, all parameters perform best when they are at the optimized value.

(d)

Fig. 10. (*continued*).

7.6. An example of event summarization

Four example of our summarization result is shown in Fig. 10, including (a) event #1 Kunming Station Massacre, (b) event #5 Ludian Earthquake, (c) event #4 Urumqi Attack, (d) event #3 Shanghai Stampede. Because of space limitation, only top 3 images and top 2 texts of each subevent are listed. We cam see that four subevents are mined from (c) event #4 Urumqi Attack. The first one is mainly about photos on crime scene, while texts introduce the plot of the event. And in subevent #2 and #3, they represent two kinds of memorial activities. Specifically, subevent #2 is the pray on the network, and #3 is the pray on the crime scene. Subevent #4 is about officers' demonstrations in Urumqi. Our method can well mine the subevents hidden in the events, and the text and image summarization offer closely related and diverse display for the events.

8. Conclusions

In this work, we proposed a novel event summarization method. Different to traditional method, we combined social user feature into multimedia co-clustering. We also proposed coarse-to-fine filtering method to eliminate noisy information and reserved relevant information for given events. Text and image summarization offer relevant and diverse information to the events. However the proposed approach is rely on the event whose name and happen time have been given. In the future work, we are committed to

make a comprehensive system that can detect breaking events, and perform text and image summarization in real time.

Acknowledgments

This work was supported in part by the NSFC, China under Grant 61732008, 61772407, 61332018 and u1531141), the National Key R&D Program of China under Grant 2017YFF0107700, the World-Class Universities (Disciplines) and the Characteristic Development Guidance Funds for the Central Universities, China, under No. PY3A022.

References

[1] J. Bian, Y. Yang, T, Chua, Multimedia summarization for trending topics in microblogs, in: Proc. CIKM, 2013, pp. 1807–1812.
[2] J. Bian, Y. Yang, H. Zhang, T. Chua, Multimedia summarization for social events in microblog stream, IEEE Trans. Multimedia (2015).
[3] X. Zhou, L. Chen, Event detection over twitter social media streams, VLDB J.-Int. J. Very Large Data Bases (2014) 381–400.
[4] K. Doman, T. Tomita, I. Ide, D. Deguchi, H. Murase, Event detection based on twitter enthusiasm degree for generating a sports highlight video, in: ACM MM, 2014, pp. 949–952.
[5] X. Qian, H. Feng, G. Zhao, T. Mei, Personalized recommendation combining user interest and social circle, IEEE Trans. Knowl. Data Eng. 26 (7) (2014) 1487–1502.
[6] A. Cui, M. Zhang, Y. Liu, S. Ma, K. Zhang, Discover breaking events with popular hashwords in twitter, in: CIKM, 2012.
[7] A.J. McMinn, Y. Moshfeghi, J.M. Jose, Building a large-scale corpus for evaluating event detection on twitter, in: CIKM, 2013.
[8] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R.M. Tripathy, S. Triukose, Spatio-temporal and events based analysis of topic popularity in twitter, in: CIKM, 2013.
[9] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proc. WWW, 2010, pp. 851–860.
[10] A. Popescu, M. Pennacchiotti, Detecting controversial events from twitter, in: CIKM, 2010.
[11] Michael Mathioudakis, Nick Koudas, TwitterMonitor: Trend detection over the twitter stream, in: SIGMOD, 2010.
[12] L. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, A. Jaimes, Sensing trending topics in twitter, IEEE Trans. Multimedia 15 (6) (2015).
[13] A. Popescu, M. Pennacchiotti, D.A. Paranjpe, Extracting events and event descriptions from twitter, in: Proc. WWW, 2011, pp. 105–106.
[14] X. Gao, J. Gao, Z. Jin, X. Li, J. Li, Extracting events and event descriptions from twitter, in: Proc. WWW, 2011, pp. 105–106.
[15] S. Prasad, P. Melville, A. Banerjee, Emerging topic detection using dictionary learning, in: CIKM, 2011.
[16] K. Chen, Y. Zhou, H. Zha, J. He, P. Shen, X. Yang, Cost-effective node monitoring for online hot event detection in sina weibo microblogging, in: WWW, 2013.
[17] W. Wang, L. He, P. Markham, H. Qi, Y. Liu, Q. Cao, L.M. Tolbert, Multiple event detection and recognition through sparse unmixing for high-resolution situational awareness in power grid, IEEE Trans. Smart Grid 5 (4) (2014).
[18] L.X. Xie, H. Sundaram, andM. Campbell, Event mining in multimedia streams, Proc. IEEE 96 (4) (2008) 623–646.
[19] M. Schinas, S. Papadopoulos, G. Petkos, Y. Kompatsiaris, P.A. Mitkas, Multimodal graph-based event detection and summarization in social media streams, in: ACM MM, 2015.
[20] O. Ozdikis, H. Oguztuzun, P. Karagoz, Evidential location estimation for events detected in twitter, in: GIR, 2013.
[21] X. Qian, Y. Zhao, J. Han, Image location estimation by salient region matching, IEEE Trans. Image Process. 24 (6) (2015) 4348–4358.
[22] Y. Zhao, X. Qian, Spatial constraint for image location estimation, in: Proc. ICMR, 2015, pp. 515–518.
[23] Y. Wu, H. Zhang, B. Xu, H. Hao, C. Liu, TR-LDA: A cascaded key-bigram extractor for microblog summarization, Int. J. Mach. Learn. Comput. 5 (3) (2015) 172–178.
[24] Y. Gong, X. Liu, Generic text summarization using relevance measure and latent semantic analysis, in: SIGIR, 2001.
[25] B. Bao, C. Xu, W. Min, M.S. Hossain, Cross-Platform emerging topic detection and elaboration from multimedia streams, ACM Trans. Multimedia Comput. Commun. Appl. 11 (4) (2015) Article 54.
[26] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, D.S. Modha, A generalized maximum entropy approach to Bregman co-clustering and matrix approximation, J. Mach. Learn. Res. (2007) 1919–1986.
[27] N. Alsaedi, P. Burnap, O. Rana, Temporal TF-IDF: A high performance approach for event summarization in twitter, in: IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, 2016, pp. 515–521.
[28] A. Weiler, M. Grossniklaus, M.H. Scholl, Run-time and task-based performance of event detection techniques for twitter, in: Advanced Information Systems Engineering, Springer International Publishing, 2015, pp. 35–49.
[29] A. Iyengar, T. Finin, A. Joshi, Content-based prediction of temporal boundaries for events in twitter, in: IEEE Third International Conference on Privacy, Security, Risk and Trust, IEEE, 2011, pp. 186–191.
[30] C.H. Lee, H.C. Yang, T.F. Chien, et al., A novel approach for event detection by mining spatio-temporal information on microblogs, in: International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2011, pp. 254–259.
[31] Y. Qu, C. Huang, P. Zhang, J. Zhang, Microblogging after a major disaster in China: A case study of the 2010 Yushu Earthquake, in: CSCW, 2011.
[32] Y. Ren, X. Qian, S. Jiang, Visual summarization for place-of-interest by social-contextual constrained geo-clustering, in: MMSP, 2015.
[33] S. Jiang, X. Qian, J. Shen, Y. Fu, T. Mei, Travel recommendation via author topic model based collaborative filtering, IEEE Trans. Multimedia 17 (6) (2015) 907–918.
[34] H. Feng, X. Qian, Recommend social network users favorite brands, in: PCM, 2013.
[35] L.D.S. Belo, Summarizing video sequence using a graph-based hierarchical approach, Neurocomputing 173 (P3) (2016) 1001–1016.
[36] M. Schinas, S. Papadopoulos, G. Petkos, et al., Multimodal Graph-based Event Detection and Summarization in Social Media Streams, 2015, pp. 189–192.
[37] K. Kumar, D.D. Shrimankar, N. Singh, Eratosthenes sieve based key-frame extraction technique for event summarization in videos, Multimedia Tools Appl. (2017) 1–22.
[38] Deepti D. Shrimankar, Navjot Singh, https://www.researchgate.net/publication/316446438_Equal_Partition_Based_Clustering_Approach_for_Event_Summarization_in_Videos, 2016/01/01, 126, http://dx.doi.org/10.1109/SITIS.2016.27.
[39] C. Kuang, J. Tang, Z. Liu, M. Sun, ImgWordle: Image and text visualization for events in microblogging services, in: Proc. AVI, 2014, pp. 371–372.
[40] A.J. McMinn, D. Tsvetkov, T. Yordanov, A. Patterson, R. Szk, J.A. Rodriguez Perez, J.M. Jose, An interactive interface for visualizing events on twitter, in: ACM SIGIR, 2014, pp. 1271–1272.
[41] R.R. Shah, A.D. Shaikh, Y. Yu, W. Geng, P. Zimmermann, G. Wu, EventBuilder: Real-time multimedia event summarization by visualizing social media, in: ACM MM, 2015.
[42] P. Wang, H. Wang, M. Liu, W. Wang, An algorithmic approach to event summarization, in: SIGMOD, 2010.
[43] Z. Wang, P. Cui, L. Xie, W. Zhu, Y. Zhu, S. Yang, Bilateral correspondence model for words-and-pictures association in multimedia-rich microblogs, in: ACM TOMM, vol. 10, 2014.
[44] X. Wang, W. Ma, G. Xue, X. Li, Multi-model similarity propagation and its application for web image retrieval, in: MM, 2004.
[45] M. Schinas, S. Papadopoulos, Y. Kompatsiaris, P.A. Mitkas, Visual event summarization on social media using topic modelling and graph-based ranking algorithms, in: ICMR, 2015.
[46] Y. Xue, X. Qian, X. Yang, Y.Y. Tang, X. Hou, T. Mei, Landmark summarization with diverse viewpoints, IEEE Trans. Circuits Syst. Video Technol. 25 (11) (2015) 1857–1869.
[47] Y. Pang, Summarizing tourist destinations by mining user-generated travelogues and photos, Comput. Vis. Image Underst. (2011) 352–363.
[48] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, D.S. Modha, A Generalized maximum entropy approach to bregman coclustering and matrix approximation, in: ACM KDD, 2004, pp. 509–514.
[49] G. Petkos, S. Papadopoulos, V. Mezaris, R. Troncy, P. Cimiano, T. Reuter, Y. Kompatsiaris, Social event dReuter, Y. Kompatsiaris, Social event detection at MediaEval: a three-year retrospect of tasks and results, in: ACM ICMR, 2014.
[50] B.K. Bao, W. Min, K. Lu, C. Xu, Social event detection with robust high-order co-clustering, in: ACM ICMR, 2013, pp. 135–142.
[51] H. Cai, Y. Yang, X. Li, Z. Huang, What are Popular: Exploring twitter features for event detection, tracking and visualization, in: MM, 2015.
[52] G. Zhao, X. Qian, X. Xie, User-Service rating prediction by exploring social users' rating behaviors, IEEE Trans. Multimed. 18 (3) (2016) 496–506.
[53] C.Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out: ACL-04 Workshop, 2004, pp. 74–81.
[54] M. Schinas1, S. Papadopoulos, Y. Kompatsiaris, MGraph: multimodal event summarization in social media using topic models and graph-based ranking, Int. J. Multimedia Inf. Retr. 5 (1) (2016) 51–69.