Multimedia Summarization for Social Events in Microblog Stream

Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua

***Abstract*—Microblogging services have revolutionized the way people exchange information. Confronted with the ever-increasing numbers of social events and the corresponding microblogs with multimedia contents, it is desirable to provide visualized summaries to help users to quickly grasp the essence of these social events for better understanding. While existing approaches mostly focus only on text-based summary, microblog summarization with multiple media types (e.g., text, image, and video) is scarcely explored. In this paper, we propose a multimedia social event summarization framework to automatically generate visualized summaries from the microblog stream of multiple media types. Specifically, the proposed framework comprises three stages, as follows. 1) A noise removal approach is first devised to eliminate potentially noisy images. An effective spectral filtering model is exploited to estimate the probability that an image is relevant to a given event. 2) A novel cross-media probabilistic model, termed *Cross-Media-LDA* (CMLDA), is proposed to jointly discover subevents from microblogs of multiple media types. The intrinsic correlations among these different media types are well explored and exploited for reinforcing the cross-media subevent discovery process. 3) Finally, based on the cross-media knowledge of all the discovered subevents, a multimedia microblog summary generation process is designed to jointly identify both representative textual and visual samples, which are further aggregated to form a holistic visualized summary. We conduct extensive experiments on two real-world microblog datasets to demonstrate the superiority of the proposed framework as compared to the state-of-the-art approaches.**

***Index Terms*—Microblog, multimedia summarization, social event.**

I. Introduction

RECENT years have witnessed the emergence of microblogging services that change the way people live, work and communicate. For example, Sina Weibo,[1] one of the largest microblogging platforms on the Web, has attracted more than 500 million registered users, and the average number of daily active users has reached 46 million by the end of 2012. The contents of microblogs are becoming more multimedia with close to 37% of Sina Weibo microblogs containing images [1]. With wide availability of information sources, rapid information propagation and ease of use, microblogging has quickly become one of the most important medium for sharing, distributing and consuming interesting contents and topics.

Fig. 1. Illustration of social events on Sina Weibo.

One of the most important functionalities of microblogging services is to monitor hot trends, also known as social events. Given the streaming data, various techniques ([2]–[4]) have been proposed for social event detection. Currently, most microblogging platforms provide the list of ongoing social events, which will offer a potentially useful service to help users to conveniently gain a quick and concise impression of the current hot social events. For example, Twitter provides the *Trends* service, and Sina Weibo provides the *Hot Topics* service (as presented in Fig. 1). However, social event detection itself does not end the story. It only provides cues of the existence of a new event, together with the tremendous volume of unorganized microblog posts, which usually offer too many details to browse. Without effective summarization mechanism, the users are often confronted with incomplete, irrelevant and duplicate information, which makes it difficult to capture the essence of the event and possible to miss information of a valuable direction. Therefore, it would be of great benefit if an effective mechanism can be provided for summarizing the detected social events. In this paper, we focus on the step after social event detection: given the microblog posts related to a detected social event, we target at mining the different divisions under this event (denoted as subevents in this paper), as well as summarizing these subevents precisely and concisely.

It is natural to formulate microblog summarization as a multi-document summarization (MDS) [5] task, which has been widely studied in information retrieval. MDS is an automatic procedure aimed at extraction of information from multiple texts about the same topic. Resulting summary report allow

Manuscript received March 30, 2014; revised September 30, 2014; accepted November 08, 2014. Date of publication December 22, 2014; date of current version January 15, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Cees G. M. Snoek.

J. Bian, H. Zhang, and T.-S. Chua are with the School of Computing, National University of Singapore, Singapore 119077 (e-mail: bian_jingwen@nus.edu.sg; hanwang@nus.edu.sg; dcscts@nus.edu.sg).

Y. Yang is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: dlyyang@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2384912

[1][Online] Available: http://www.weibo.com

readers to quickly capture the essential information contained in a large cluster of documents. Most of the previous MDS techniques are designed for well-organized texts, such as news articles. However, two challenges make it greatly difficult to directly apply the traditional MDS techniques to summarize microblogs: 1) word restriction (e.g., a maximum 140 characters for Twitter[2]) and 2) unreliable microblog contents. Recently, several attempts ([6]–[8]) have been made to summarize microblog text by considering the above limitations. Nonetheless, the performance is still far from satisfactory because they did not address the problem of lacking descriptive power caused by word restriction and noisy content.

Different from traditional documents that contain only textual objects, microblogs are comprised of contents of various media types, such as images and video links. For instance, of the 310,097 microblogs in our *Social Trends* dataset (refer to Section V for details about the dataset), 114,426 (36.9%) contain images. Such high proportion of multimedia contents are potentially precious resources for handling the different problems ([9], [10]). The benefit of incorporating different media types into summarization is three-fold: 1) In many cases, images contain essential information which could not be completely expressed by the microblog texts. Therefore, the visual information is of great significance for summarizing the event and remedying the descriptive power of short texts. In addition, when the emphasis of an event lies in the visual part, it will not be meaningful if only textual summary is generated. 2) Multimedia contents can facilitate subevent discovery. Intuitively, given a social event, multimedia contents from different subevents should have lower visual similarity while those within the same subevent should have higher visual similarity. Thus, discriminative information embedded in visual information of multimedia contents can be exploited as critical cues for subevent discovery. 3) Incorporating concrete multimedia exemplars into summarization can assist users to gain a more visualized understanding of interesting events.

It is non-trivial to integrate textual and visual information to generate comprehensive summaries in the circumstance of microblogs. If the intrinsic correlations between textual and visual information are not well explored, they may exert negative influence on each other. In our previous work [11], we target at this problem, and proposed a framework to automatically generate multimedia summary for trending topic in microblogs. By comparing with many state-of-the-art summarization methods, we proved the superiority of our proposed framework in terms of effectiveness and practicality. However, in the mean time, we also discovered another severe problem regarding to the quality of microblog posts. As the input of our problem is a stream of microblog posts related to a detected social event, and the detection process is usually text-based with no visual information taken into consideration, therefore, the inconsistency of textual part and visual part in the same microblog is very common, we are faced with irrelevant microblog texts and/or images in the input data stream. For instance, according to our statistics, the percentage of relevant images in our datasets is only 67.1% (refer to Section V for details). Directly utilizing such noisy images may severely degrade the performance of subevent discovery and summarization. Taken this issue into consideration, we extend the previous work by adding an important component into our framework, which tackles the problem of removing irrelevant data before the succeeding procedures of subevent discovery and summary generation.

To sum up, in this paper, we propose a novel multimedia social event summarization framework to generate holistic visualized summary from the microblogs with multiple media types. Specifically, the proposed framework comprises three stages: removal of irrelevant data, cross-media subevent discovery and multimedia summary generation. First, we devise a data cleansing approach to automatically eliminate those irrelevant/noisy images. An effective spectral filtering model is exploited to estimate the probability that an image is relevant to a given event. In the second stage, we propose a novel cross-media probabilistic model, termed *Cross-Media-LDA* (CMLDA), to jointly exploit the microblogs of multiple media types for discovering subevents. The CMLDA model not merely well explores and exploits the intrinsic correlations among different media types, but also simultaneously characterizes both the general distribution and the subevent specific distribution from the microblog data of various media types for reinforcing the subevent discovery process. Besides, this step could also handle the noise of the input data, and remove those microblog examples from the next summarization step. Finally, based on the cross-media distribution knowledge of all the discovered subevents, we generate a holistic visualized summary for the social events by pinpointing both the representative textual and visual samples in a joint fashion. In particular, by utilizing the cross-media distributions of microblog text, we specify three criteria, namely coverage, significance and diversity to measure the summarization capability of individual textual samples. We then devise a greedy algorithm for identifying the representative microblog texts based on the combination of the three criteria. For visual summarization, we employ the cross-media knowledge of the subevents as the prior knowledge for ranking the visual samples and selecting the most representative ones. In order to improve the descriptive power and the diversity of viewpoints, we first partition the images within a subevent into groups via spectral clustering. Then, for each group we apply a manifold algorithm with the cross-media prior knowledge as initial ranking scores to identify the top-ranked image as representative. It is remarkable that both the textual and visual summarization processes utilize the cross-media knowledge of the discovered subevents and thus are intrinsically connected to reinforce each other.

The rest of the paper is organized as follows. Related works are briefly summarized and discussed in Section II. Section III introduces problem formulation and framework overview. We elaborate the details of the proposed multimedia social event summarization framework in Section IV, including noise removal, subevent discovery and microblog summary generation. Experimental results are reported and discussed in Section V, followed by the conclusion in Section VI.

II. Related Work

Multi-document summarization has drawn much attention in the past two decades. General MDS method can be separated into two types: extractive summarization and abstractive summarization. The former one usually ranks the sentences in the docu-

[2][Online] Available: http://twitter.com

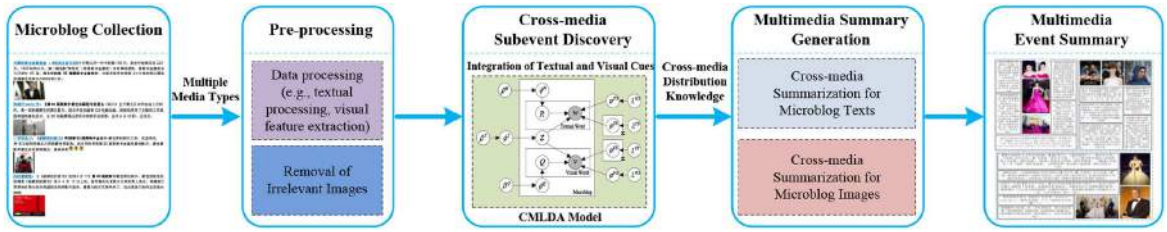


Fig. 2. Flowchart of the proposed multimedia social event summarization framework.

ments according to their salient scores calculated by a set of predefined features and then extracts top ranked sentences; while the latter involves information fusion, sentence compression and reformulation. Over the years, many methods have been proposed for MDS, and most of them are extractive methods. In this work, we focus on extractive summarization, and take each microblog as the extractive unit instead of a single sentence.

Notable MDS methods include SumBasic [12] and centroid-based algorithm. The underlying premise under SumBasic is that words which occur more frequently across documents have a higher probability of being selected for human created multi-document summaries than words that occur less frequently. MEAD [13], one implementation of centroid-based method, discovers the centroid of each document cluster, and extracts sentences closest to the centroid. Maximal marginal relevance (MMR) [14] method is adopted to remove redundancy sentences. Another direction is to use graph-based methods to rank sentences based on the vote between each other. TextRank [15] and LexPageRank [16] use algorithms like PageRank and HITS to compute sentence importance. These methods first construct a graph representing the relationship between sentences and then evaluate the importance of each sentence based on the topology of the graph. Some other methods are designed to identify semantically important sentences for summary generation. For example, [17] uses the latent semantic analysis (LSA) technique to select highly ranked sentences; and a hierarchical LDA-style model is utilized in [18] to represent content specificity as a hierarchy of topic vocabulary distributions and then sentences are selected according to these distributions. Other methods include NMF-based topic-specific summarization [19]; conditional random fields (CRF) based summarization [20]; and hidden Markov model (HMM) based method [21]. Most recently, Wang *et al.* [22] proposed a framework to summarize multiple documents via sentence-level semantic analysis and symmetric matrix factorization. While all these methods are designed for well-organized formal texts, directly applying them on microblog dataset is not very suitable.

With the development of microblogging, many works have shifted their focus to process microblog data. Most of the prior work on Twitter data summarization are about topic-level summarization. Sharifi *et al.* [8] summarized Twitter hot topics through finding the most commonly used phrase that encompasses the topic phrase. In [6], clustering algorithms are introduced for Twitter topic summarization to select multiple posts that convey information about a given topic without being redundant. Chakrabarti *et al.* [7] formalized the problem of tweets summarizing for some highly structured and recurring events, and offered a solution based on learning the underlying hidden state representation of the event via HMM. Lin *et al.* [23] generated event storyline of an ongoing event via graph optimization for microblogs. The temporal information is utilized for event representation, and this framework is only suitable for relatively long-term events. However, the hot period of most social events to be summarized for our task is usually one day only, which makes the temporal information less valuable for such short term. One problem for all approaches above is that they only focused on the textual information, while the precious multimedia resources are neglected.

Several previous works have leveraged the importance of multimedia resources [24], [25] and proposed methods for multimodal information representation. Yan *et al.* exploited the fact that news documents are often accompanied by pictures and proposed a graph-based framework to generate a timeline summary for a news topic in [26]. Mutual Summarization is proposed in [27] to summarize images by text and visualize text utilizing images. The target domain for these previous efforts is news. Compared with user-generated microblog contents, news articles are formalized and unified across multiple documents. Besides, the number of corresponding images in news articles is usually small, and the images are of high quality and high relevance. Hence it is not possible to adopt multimedia summarization techniques in news domain to microblog data.

III. Framework Overview

Suppose we are given a microblog stream $\mathcal{M} = \{M_1, \ldots, M_{|\mathcal{M}|}\}$ related to the same social event $\mathcal{E}$, which can be either provided by any social event service of online microblog platform or detected by any social event detection method. Each microblog $M_i = \{T_i, I_i\}$ consists of two components: textual component T_i and visual component I_i. Note that I_i may be empty, which means M_i contains no visual sample. $|\cdot|$ denotes the cardinality of a set. The objective of our framework is to automatically generate a multimedia summary (i.e., both textual and visual) from the microblog collection $\mathcal{M}$ for revealing multiple subevents of the event $\mathcal{E}$. For event $\mathcal{E}$, we define its event-level summary as the union of all its subevents' summaries. For each subevent, a subevent-level summary comprises both textual and visual exemplars selected from $\mathcal{M}$. The flowchart of the proposed multimedia social event summarization framework is illustrated in Fig. 2. As can be seen, there are three main stages in the whole process. In the first stage, we preprocess both the textual and visual data of microblogs as well as eliminate irrelevant images, which results in a more reliable microblog collection. In the second stage, we discover subevents from microblogs by using the proposed CMLDA model. This model determines the subevent assignment for each microblog, as well as the cross-media distribution knowledge for all subevents. Then, the next stage substantially exploits the cross-media knowledge of each individual subevent for jointly identifying both the representative

textual and visual exemplars, which forms the subevent-level summary. Finally, we aggregate all the subevent-level summaries to derive the holistic summary for the social event.

IV. Multimedia Social Event Summarization

In this section, we elaborate the details of the proposed multimedia social event summarization framework, including the removal of irrelevant data, the cross-media subevent discovery and the multimedia summary generation.

A. Removal of Irrelevant Data

As a kind of user-generated content, the quality of microblogs cannot be guaranteed. It has been observed that many microblog images are irrelevant to their corresponding texts (e.g., spam images). Directly applying our framework on such noisy image set may severely degrade the performance of the summarization. Since the input microblog collection is gathered with text-based methods, the problem of noisy images is more severe than that of noisy texts. Therefore, it is necessary to first pre-filter microblog images to eliminate those noisy images. For the problem of noisy texts, it will be addressed in the following subevent discovery procedure.

Specifically, we develop the noise removal procedure by exploiting a spectral filtering model [28]. Without loss of generality, suppose we have n microblog images $X = \{x_1, x_2, \ldots, x_n\}$ corresponding to all the non-empty images I_i of the given social event $\mathcal{E}$, where $x_i \in \mathbb{R}^d$ and d is the dimension of visual space. We first build a neighborhood graph $G = (V, E, W)$, where V is a vertex set composed of n vertices representing our n images in X, $E \subseteq V \times V$ is an edge set connecting neighboring vertices, and $W \in \mathbb{R}^{n\times n}$ is a weighting matrix measuring the strength of the edges, i.e., the similarity between two data points. There are various methods to compute W. In this work, we adopt the widely used k-Nearest-Neighbors similarity graph

$$W_{ij} = \begin{cases} \exp\left(-\frac{d(x_i,x_j)^2}{\sigma^2}\right), & \text{if } x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i) \\ 0, & \text{otherwise} \end{cases}$$

where $d(\cdot,\cdot)$ is a distance measure such as Euclidean distance, and σ is the bandwidth parameter. $\mathcal{N}_k(x_i)$ denotes the set of k nearest neighbors of x_i in X. We further define D as a degree matrix whose diagonal elements $D_{ii} = \sum_{j=1}^{n} W_{ij}$. With D and W, the normalized graph Laplacian is defined as

$$L = I - D^{-1/2} W D^{-1/2}.$$

As previously discussed, an intuition is that images depicting the same subevent should be visually similar to each other. Therefore, it is reasonable to assume that the truely relevant images should reside in multiple high-density regions, while the irrelevant images will present a more random distribution. It has been demonstrated [29] that when data points have formed clusters, each high density region implicitly corresponds to certain low-frequency (smooth) eigenvector. The data points which belong to the region will take relatively large absolute values corresponding to the eigenvector, while for data points elsewhere, the values are close to zero. With this assumption, we can exploit the spectrum of the k NN similarity graph G, which is a set of eigenvalue/eigenvector pairs $\{\lambda_i, \mathbf{u}_i\}_{\mathrm{i}=1}^{\mathbf{n}}$ of the normalized graph Laplacian L to find the high density regions. For simplicity, we assume that the eigenvalues are sorted in a nondecreasing order, thus the top eigenvectors have the lowest frequency.

Let $\mathbf{y} \in \mathbb{R}^{n\times 1}$ be a label vector indicating the relevance of each image to the given event. Ideally, $\mathbf{y}$ takes the value of 1 for all relevant images and 0 for noisy ones. Consider the top m smoothest eigenvectors $\mathbf{u}_2, \ldots, \mathbf{u}_{\mathbf{m}+1}$ as eigenbases ($\mathbf{u}_1$ is eliminated because it is nearly constant and $\lambda_1 = 0$ when the graph is connected, thus does not form any region). According to the multi-region assumption, $\mathbf{y}$ should lie in the subspace spanned by these eigenbases. Let $U = [\mathbf{u}_2, \ldots, \mathbf{u}_{\mathrm{m}+1}] \in \mathbb{R}^{\mathbf{n}\times\mathbf{m}}$, $\Lambda = diag(\lambda_2, \lambda_3, \ldots \lambda_{m+1})$ and $\mathbf{y} = 1$ initially. The spectral filter reconstructs the noisy label vector $\mathbf{y}$ with the sparse eigenbases U by solving the following problem:

$$\min_{\mathbf{z}\in\mathbb{R}^{\mathbf{m}}} \|U\mathbf{z} - \mathbf{y}\|^2 + \alpha_1 \|\mathbf{z}\|_1 + \alpha_2 \mathbf{z}^{\mathbf{T}} \Lambda \mathbf{z} \tag{1}$$

where $\mathbf{z}$ is the sparse coefficient vector, $\|\mathbf{z}\|_1 = \sum_{j=1}^{m} |z_j|$ is the ℓ_1-norm. α_1 and α_2 are two regularization parameters. Note that the last term $\mathbf{z}^T\Lambda\mathbf{z} = \sum_{i=1}^{m} \lambda_{i+1} {z_i}^2$, which is actually a weighted ℓ_2-norm, imposes that smoother eigenbases with smaller eigenvalues are preferred in the reconstruction of $\mathbf{y}$.

Once the solution $\mathbf{z}$ to Eq. (1) is obtained, the truly relevant label vector $\hat{\mathbf{y}}$ is set to $round(U\mathbf{z})$, where the function $round(\cdot)$ is defined as follows:

$$(round(\mathrm{x}))_{\mathrm{i}} = \begin{cases} 1, & x_i \geq \theta \cdot \max\{\mathrm{x}\} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where θ is the threshold indicating the confidence level of a data to be regarded as relevant. With the final label vector $\hat{\mathbf{y}}$, we eliminate those image samples with $\hat{y}_i = 0$, and retain a more reliable image set.

B. Cross-Media Subevent Discovery

In this subsection, we propose a novel cross-media probabilistic model, termed *Cross-Media-LDA* (CMLDA), to discover subevents by jointly exploring the intrinsic correlation between the textual and visual aspects of microblogs. The CMLDA model substantially characterizes the multiple facets of the social event by exploring two underlying properties of various media types, i.e., *inter-media consistency* and *intra-media discrimination*. Besides, this model is also capable of eliminating possible noisy microblog posts from the data collection gathered for the following summarization process.

Inter-Media Consistency: It is observed that the microblogs associated with the same social event contain various inter-correlated media types, such as texts and images. If we can properly capture and model the intrinsic correlations among these media types, we may achieve a better understanding of the social event. Intuitively, different media types of the same event should be related to certain common topics or share some common high-level semantics. In other words, the semantics should be consistent across different media types. Based on this analysis, we model the common semantics shared among different media types via a subevent indicator Z, which is able to jointly generate both the textual words and visual words in the microblogs. With the cross-media subevent indicator, we manage to capture the inter-media correlations for effective subevent discovery. It is worth noting that while the traditional

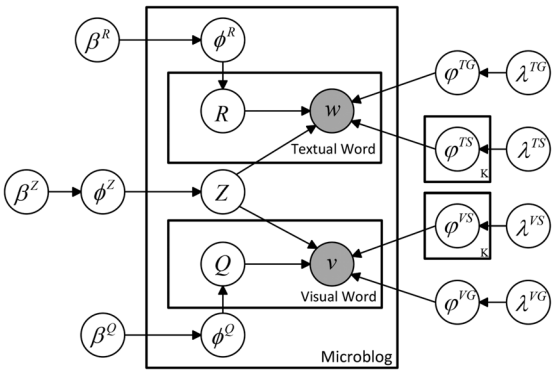


Fig. 3. Graphical model representation of the CMLDA model.

latent dirichlet allocation (LDA) [30] model assigns multiple topics to each individual document and one topic for each word, the proposed CMLDA model is designed to associate only one topic (subevent) with each individual microblog. The underlying reason is that microblog content is usually short and focused, and thus it is reasonable to assume that each microblog is related to only one subevent.

Intra-Media Discrimination: Within each individual media type, it is non-trivial to directly employ traditional topic modeling approach (e.g., LDA) to discover the subevents within the same social event because the semantics of different subevents may be heavily overlapped, while we target at discovering the discriminative knowledge of each subevent.

Normally, we may assume that all subevents of the same event share certain general words indicating common semantics related to the social event; while each individual subevent uniquely possesses certain specific semantics, which distinguishes itself from other subevents. Take "Lushan Earthquake" as an example, "earthquake", "Lushan" and "death" are more likely to be general words; while words like "hypocenter", "collapse" and "Premier" are more probable to appear in different subevents. If the proportion of general contents is large, then they may dominate the result. In order to exclude the influence of general contents and discover discriminative cues for each subevent, two new latent variables R and Q are introduced to guarantee *intra-media discrimination* in the generation of textual and visual words, respectively. For each textual (visual) word, R (Q) indicates whether it is generated from the general distribution or from the specific distribution corresponding to its subevent.

CMLDA Modeling and Inference: In this part, we elaborate the details of the modeling and inference processes of the CMLDA model. Fig. 3 illustrates the graphical model representation, and the key notations are listed in Table I. The generation process is as follows.

1. For the event $\mathcal{E}$, draw $\varphi^{TG} \sim Dir(\lambda^{TG})$ and $\varphi^{VG} \sim Dir(\lambda^{VG})$, indicating the general textual and visual distribution, respectively. Then draw $\phi^Z \sim Dir(\beta^Z)$, which indicates the distribution of subevents over the microblog collection corresponding to $\mathcal{E}$.
2. For each subevent, draw $\varphi_k^{TS} \sim Dir(\lambda^{TS})$ and $\varphi_k^{VS} \sim Dir(\lambda^{VS})$, $k \in \{1, 2, \ldots, K\}$, corresponding to the specific textual and visual distribution.
3. For each microblog M_i, draw $Z_i \sim Multi(\phi^Z)$, corresponding to the subevent assignment for M_i. Then draw $\phi_i^R \sim Dir(\beta^R)$, indicating the general/specific textual word distribution of M_i. Similarly, draw $\phi_i^Q \sim Dir(\beta^Q)$.
4. For each textual word position of M_i, draw a variable $R_{ij} \sim Multi(\phi_i^R)$:
   - If R_{ij} indicates *General (G for short)*, then draw a word $w_{ij} \sim Multi(\varphi^{TG})$.
   - If R_{ij} indicates *Specific (S for short)*, then draw a word w_{ij} from the Z_i-th specific distribution $w_{ij} \sim Multi(\varphi_{Z_i}^{TS})$.
5. The generation of visual words is similar to step 4.

TABLE I
LIST OF KEY NOTATIONS

Notations	Descriptions
$\mathcal{E}$	Event indicator
ϕ^Z	Subevent distribution for event $\mathcal{E}$
Z	Subevent indicator
φ^{TG}, φ^{VG}	General textual word and visual word distribution
φ^{TS}, φ^{VS}	Specific textual word and visual word distribution
ϕ^R, ϕ^Q	Distribution of textual and visual word belonging to general or specific part
R, Q	Indicator to determine the textual (R) or visual (Q) word is generated from the general distribution or from the corresponding specific subevent distribution
w, v	Textual and visual word in a microblog

In the CMLDA model, the subevent indicator Z as well as the general/specific indicator R and Q are latent variables to be inferred from observations, i.e., textual and visual words. We use Gibbs sampling to achieve the inference due to its efficiency and effectiveness in handling high-dimensional data. The update rules for latent variables are shown as follows:

$$P(R_{ij} = S|Z_i, w_{ij}, \cdots) \propto \frac{N_{i,\neg j}^R(S) + \beta_S^R}{N_{i,\neg j}^R + \beta_G^R + \beta_S^R} \times \frac{N_{Z_i,\neg j}^{TS}(w_{ij}) + \lambda^{TS}}{\sum_{t \in V^t}\left(N_{Z_i,\neg j}^{TS}(t) + \lambda^{TS}\right)}$$

$$P(R_{ij} = G|w_{ij}, \cdots) \propto \frac{N_{i,\neg j}^R(G) + \beta_G^R}{N_{i,\neg j}^R + \beta_G^R + \beta_S^R} \times \frac{N_{\neg j}^{TG}(w_{ij}) + \lambda^{TG}}{\sum_{t \in V^t}\left(N_{\neg j}^{TG}(t) + \lambda^{TG}\right)}$$

$$P(Q_{ij} = S|Z_i, v_{ij}, \cdots) \propto \frac{N_{i,\neg j}^Q(S) + \beta_S^Q}{N_{i,\neg j}^Q + \beta_G^Q + \beta_S^Q} \times \frac{N_{Z_i,\neg j}^{VS}(v_{ij}) + \lambda^{VS}}{\sum_{u \in V^v}\left(N_{Z_i,\neg j}^{VS}(u) + \lambda^{VS}\right)}$$

$$P(Q_{ij} = G|v_{ij}, \cdots) \propto \frac{N_{i,\neg j}^Q(G) + \beta_S^Q}{N_{i,\neg j}^Q + \beta_G^Q + \beta_S^Q} \times \frac{N_{Z_i,\neg j}^{VG}(v_{ij}) + \lambda^{VG}}{\sum_{u \in V^v}\left(N_{Z_i,\neg j}^{VG}(u) + \lambda^{VG}\right)}$$

$$P(Z_i = k|R_{i\cdot}, Q_{i\cdot}, w_{i\cdot}, v_{i\cdot}, \cdots) \propto \frac{N_{\neg i}^Z(k) + \beta^Z}{\sum_{l=1}^K\left(N_{\neg i}^Z(l) + \beta^Z\right)} \times \prod_{Q_{ij}=S} \frac{N_{k,\neg i}^V(v_{ij}) + \lambda^V}{\sum_{u \in V^v}\left(N_{k,\neg i}^V(u) + \lambda^V\right)} \times \prod_{R_{ij}=S} \frac{N_{k,\neg i}^S(w_{ij}) + \lambda^S}{\sum_{t \in V^t}\left(N_{k,\neg i}^S(t) + \lambda^S\right)}$$

where V^t and V^v denote the textual and visual vocabulary, respectively. The variables with subscript i are corresponding to the i-th microblog M_i, while subscript j correspond to the j-th textual/visual word. $N(\cdot)$ stores the number of samples satisfying certain requirements during the iterative sampling process. For example, $N^{TS}_{k,\neg i}(t)$ represents the number of word t (excluding the words in M_i) in the k-th specific textual distribution.

After Gibbs sampling, we obtain the latent variables. Besides, K specific distributions φ^{TS} and φ^{VS} can also be easily computed. For a textual word $w \in V^t$, $\varphi^{TS}_k(w)$ measures the probability of w appearing in the k-th specific distribution. It is similar for visual distribution $\varphi^{VS}_k(u)$. Therefore, they can be evaluated as follows:

$$\varphi^{TS}_k(w) = \frac{N^w(Z=k,R=S)+\lambda^{TS}}{\sum_{t\in V^t}(N^t(Z=k,R=S)+\lambda^{TS})} \quad (3)$$

$$\varphi^{VS}_k(u) = \frac{N^u(Z=k,Q=S)+\lambda^{VS}}{\sum_{u\in V^v}(N^u(Z=k,Q=S)+\lambda^{VS})}. \quad (4)$$

With the CMLDA model, textual and visual components will facilitate each other to discovery the cross-media knowledge of the subevents hidden in the event. The obtained textual/visual distribution pair (φ^{TS}, φ^{VS}) depicts the discriminative multimedia cues for each subevent. According to the subevent indicator Z for each microblog, the CMLDA model partitions the microblog collection $\mathcal{M}$ into K subsets $\{\mathcal{S}_k\}|_{k=1}^K$ corresponding to K subevents where each subset contains both textual part $\mathcal{S}^t_k$ and visual part $\mathcal{S}^v_k$. Intuitively, if a subevent contains a small number of textual or visual samples, the topic of this subevent may not be important or related to the event. We argue that such subevents are probably composed of those noisy microblogs and should be removed. In our work, we remove all subsets whose sizes are smaller than $\epsilon \times |\mathcal{M}|$, where ϵ is the threshold. In the following subsection, we will employ the cross-media knowledge achieved with CMLDA for the summarization.

C. Multimedia Summary Generation

In this subsection, we explore how to utilize the cross-media distribution knowledge of all the discovered subevents to facilitate the generation of the holistic visualized summary with various media types for social events.

Cross-Media Summarization for Microblog Texts: In this part, we propose a method for text summarization based on the cross-media distribution information inferred from both the textual and visual aspects of the microblogs in the subevent discovery procedure. Specifically, a greedy algorithm is developed to sequentially select representative samples based on a novel selection criterion, which takes three fundamental requirements into consideration:

Coverage: Intuitively, if a summary is able to well "cover" the information of its corresponding subevent, then the word distributions over both of them should be close to each other. We use the similarity of word distributions over a summary and its corresponding subevent for measuring coverage. Denote $\mathcal{G}_k$ as the current summary set consisting of the selected samples, then the word distribution over $\mathcal{G}_k$, denoted as $\Theta_{\mathcal{G}_k}$, can be estimated as

$$p(w|\Theta_{\mathcal{G}_k}) = \frac{tf(w,\Theta_{\mathcal{G}_k})}{\sum_{t\in V^t} tf(t,\Theta_{\mathcal{G}_k})} \qquad \forall w \in V^t \quad (5)$$

where $tf(w,\Theta_{\mathcal{G}_k})$ denotes the term frequency of word w in $\mathcal{G}_k$. We use φ^{TS}_k as the word distribution over the corresponding subevent, which is the distribution estimated in the learning process of CMLDA model [Eq. (3)]. We employ Kullback-Leibler (KL) divergence to measure the distance of two distributions D_1 and D_2

$$D_{KL}(D_1\|D_2) = \sum\nolimits_w p(w|D_1)\log\frac{p(w|D_1)}{p(w|D_2)}. \quad (6)$$

Given the current summary set $\mathcal{G}_k$, the new sample T_i to be selected should be the one which makes the new summary (i.e., $\mathcal{G}_k \cup \{T_i\}$) achieve the best coverage (i.e., minimize the distance between $\Theta_{\mathcal{G}_k\cup\{T_i\}}$ and φ^S_k). Therefore, the coverage of each candidate T_i could be measured by the following criterion:

$$\mathcal{U}_C(T_i) = D_{KL}(\Theta_{\mathcal{G}_k\cup\{T_i\}}\|\varphi^{TS}_k). \quad (7)$$

Significance: In the circumstance of microblogging, each microblog can propagate between users by the repost action. In general, the popularity of a microblog can be revealed from the repost number. A large repost number implies that the microblog attracts a lot of attention and interest from other users, and hence indirectly represents the importance of this microblog. The users will be more satisfied if more of these hot microblogs are shown in the summary. Therefore, we use a smooth function over the repost number to measure the significance of a candidate

$$\mathcal{U}_S(T_i) = \log(\mathrm{RepostNum}(T_i)+1). \quad (8)$$

Diversity: The information diversification is favored in the final summary. We take the information redundancy into consideration for sample selection. Consider a candidate T_i, the redundancy it brings to the summary set can be measured by the similarity between this candidate and the previously generated summary, which is

$$\mathcal{U}_D(T_i) = D_{KL}(\Theta_{T_i}\|\Theta_{\mathcal{G}_k}). \quad (9)$$

Overall Selection Score: The overall selection score is defined as a weighted linear combination of the scores of coverage, significance and diversity. Since small distance $\mathcal{U}_C(T_i)$ indicates high coverage, we compute the overall selection score as

$$\mathcal{U}(T_i) = \omega_1(1-\mathcal{F}(\mathcal{U}_C(T_i))) + \omega_2\mathcal{F}(\mathcal{U}_S(T_i)) + \omega_3\mathcal{F}(\mathcal{U}_D(T_i))$$

where $\omega_1, \omega_2, \omega_3$ are trade-off parameters with $\sum_i \omega_i = 1$. $\mathcal{F}(x) = 1/(1+exp(-x))$ is a logistic increasing function for normalizing all the scores to the interval $[0,1]$.

With the above selection score for all the microblog samples, we may derive a greedy algorithm for representative sample selection. In each iteration, we select the one with the largest score from all the remaining samples.

Cross-Media Summarization for Microblog Images: Consider the visual subset $\mathcal{S}^v_k$, which contains all images related to the k-th subevent. The objective of this step is to employ the cross-media knowledge of the discovered subevents to reinforce the selection of the most representative image samples. The selected images should provide enough visually descriptive power as well as diverse viewpoints. We develop a two-step approach to automatically select representative images satisfying the above two criteria. We first partition the images within a subevent into groups via spectral clustering.

Then, for each group we apply a manifold algorithm with the cross-media prior knowledge as initial ranking scores to identify the top-ranked image as representative.

Clustering Step: With the similarity matrix W previously constructed in the step of the noise removal, the similarity matrix for $\mathcal{S}_k^v$ can be directly obtained by extracting the columns and rows corresponding to images in $\mathcal{S}_k^v$, i.e., $W^k = [W_{ij}|I_i, I_j \in \mathcal{S}_k^v]$. Then normalized cut is applied to the image set, and visual diversity is achieved across clusters.

Ranking Step: In order to discover images with best representative ability within each cluster, we adopt manifold ranking algorithm to rank the images. Let $\mathbf{r}$ denote the vector of ranking score, manifold ranking defines an iterative update process as follows:

$$\mathbf{r}^{t+1} = \gamma \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}\mathbf{r}^{t} + (1-\gamma)\mathbf{h} \tag{10}$$

where $\mathbf{h}$ represents the vector of initial ranking scores, which is an all-one vector in standard manifold ranking setting. However, in our scenario, we expect $\mathbf{h}$ to possess the prior knowledge of the importance of each image. Recall that with CMLDA model, we have achieved the discriminative visual information for this subevent, which is φ_k^{VS}. Intuitively, if an image is more consistent with φ_k^{VS}, it would have better descriptive ability for the whole subevent image set, and should gain more emphasis. Therefore, instead of all-one vector which takes equal weighting for all images, we express $\mathbf{h}$ as prior knowledge measured by the KL divergency of an image I_i and φ_k^{VS}, i.e., $h_i = 1 - \mathcal{F}(D_{KL}(I_i \| \varphi_k^{VS}))$. By integrating the prior knowledge in the ranking scheme, the descriptive ability for the cluster as well as for the subevent image set are both taken into consideration. Note that $\mathbf{r}$ has a closed form when the update process converges

$$\mathbf{r} = (1-\gamma)(\mathbf{I} - \gamma \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2})^{-1}\mathbf{h} \tag{11}$$

Finally, the image with the largest ranking value in $\mathbf{r}$ is selected from each cluster to construct the visual summarization set.

V. Experiments

A. Dataset and Experimental Settings

We conducted the evaluation of our framework on two datasets that were collected by ourselves: (1) *Social Trends*, which include 20 trending topics that were listed as hot trends in February 2013 by *Sina Weibo*. For each trending topic, we crawled the related microblogs in the life cycle of this topic using the trending topic API provided by Sina Weibo. The total number of microblogs is 310,097, of which 114,426 contain image; (2) *Product Events*, which was collected by Gao *et al.* [31]. It includes 20 product-related events and 13,932 microblogs, and 11,736 of them contain image. The detailed event lists of both datasets are shown in the appendix. Due to limited information appended to repost action, only the original microblogs are included in our datasets. In order to evaluate the quality of the generated summaries, five volunteers were invited to manually generate a textual summary for each event as golden standard individually. Each manually generated summary consists of 50 microblogs selected from the microblog datasets.

In text pre-processing procedure, we first segmented Chinese words using IKAnalyzer,[3] and then removed the stop words, low-frequency words with document frequency of less than 5 and mentions (@somebody) from textual vocabulary. Texts containing less than 3 words were also eliminated. For visual feature extraction, scale-invariant feature transform (SIFT) descriptors were first extracted from each image. Then we trained a codebook of 1,000 visual words with descriptors sampled from images of all events. With the trained codebook, each descriptor was quantized into a visual word. Each image was further represented as a 1,000-dimensional ℓ_2-normalized bag-of-visual-words feature vector.

[3][Online] Available: http://code.google.com/p/ik-analyzer

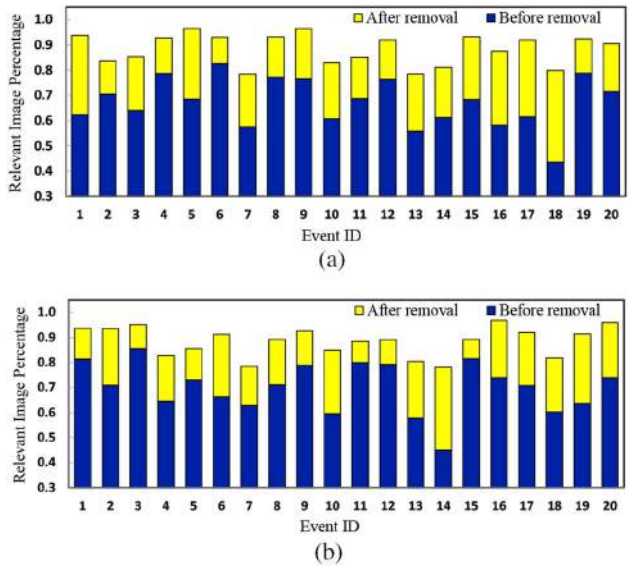


Fig. 4. Effects of irrelevant image removal. (a) *Social Trends*. (b) *Product Events*.

When constructing the image similarity graph, we set the number of nearest neighbors k to 20 and bandwidth parameter σ to 0.1. For the spectral filtering model in Eq. (1), we use $m = 40$ eigenbases for label vector reconstruction, and set $\alpha_1 = 0.1$, $\alpha_2 = 0.1$. For concentration parameters in CMLDA model, as stated in [18], the more specific a distribution is meant to be, the smaller its parameter. Accordingly, we set $\lambda^{TG} = 0.1$, $\lambda^{TS} = 0.01$, $\lambda^{VG} = 1$, $\lambda^{VS} = 0.1$, $\beta^R = 0.1$, $\beta^Q = 0.1$, and $\beta^Z = 1$. For the final representation image selection procedure, the parameter γ is set to 0.85. The threshold ϵ is set to $0.3 \times 1/K$ and all subevents with size smaller than $\epsilon \times |\mathcal{M}|$ are removed. The total number of the selected microblogs is chosen to be 50, which is the same as the number of microblogs in the gold standards. The 50 microblogs quota are assigned to the remaining subevents according to the proportion of microblog number in each subevent.

B. Capability of Irrelevant Image Removal

We demonstrate the capability of our developed irrelevant image removal component in Fig. 4. The percentage of relevant images before and after the removal procedure for each event is listed. As aforementioned, the original image collections of all the social events contain many images that are irrelevant to the corresponding event. The average percentage of relevant images is 67.1% and 69.9%, respectively, across all events for the two datasets. We apply spectral filtering on the image collection of each individual event separately. As shown in Eq. (2), one important factor which controls the performance of spectral filtering is the parameter θ of the $round(\cdot)$ function. There is a

Fig. 5. Illustrative examples of removed images and those remained after irrelevant image removal.

tradeoff between the performance of irrelevant image removal and the number of remaining images: in general, the higher the relevance percentage, the smaller the number of remaining images. In our framework, the quality of image collections is very crucial for the cross-media subtopic discovery and summarization. In our experiments, the controlling parameter θ is set to 0.5 for *Social Trends* dataset, which results in a relatively high relevance percentage (88.4%), as well as a reasonable number of images (54,800 images, or 51% of the original collection size). Similarly, we set θ to 0.6 for *Product Events*, and achieved 6,570 remaining images with 91.5% of them being relevant. On average, our proposed method improves the percentage of relevant images by around 21%. Fig. 5 shows several examples of removed and remaining images for Event #1 of *Social Trends* and Event #13 of *Product Events*. As can be clearly seen, for both events, the exemplars with high rank orders are truly relevant to the corresponding events while those images with low ranks are really noisy.

C. Summarization Performance

In this subsection, we evaluate the effectiveness of our proposed framework as compared to several summarization approaches. For fairness of evaluation, we select 50 microblogs for all the comparing approaches to form the summaries. For evaluation metric, we employ ROUGE evaluation toolkit [32] which automatically determines the quality of a summary as compared to human generated golden standards. In particular, F-measure scores of ROUGE-1, ROUGE-2, ROUGE-W (with W set to 1.2) and ROUGE-SU are reported. Take ROUGE-N as an example. Denote the golden standards as GS, and the generated summary as S, ROUGE-N-Recall is an N-gram recall metric computed as follows:

$$\text{ROUGE-N-Recall} = \frac{\sum_{I \in GS} \sum_{\text{N-gram} \in I} \text{Count}_{\text{match}}(\text{N-gram})}{\sum_{I \in GS} \sum_{\text{N-gram} \in I} \text{Count}(\text{N-gram})}$$

and ROUGE-N-Precision is an N-gram precision matric as follows:

$$\text{ROUGE-N-Precision} = \frac{\sum_{I \in S} \sum_{\text{N-gram} \in I} \text{Count}_{\text{match}}(\text{N-gram})}{\sum_{I \in S} \sum_{\text{N-gram} \in I} \text{Count}(\text{N-gram})}$$

For the ROUGE-N value reported in our experimental results, we adopt the F score of the above recalled based and precision based metrics

$$\text{ROUGE-N} = \frac{2 \times \text{ROUGE-N-Precision} \times \text{ROUGE-N-Recall}}{\text{ROUGE-N-Precision} + \text{ROUGE-N-Recall}}.$$

We compare our proposal with the following multi-document summarization approaches.

- RANDOM: which selects all samples randomly.
- LSA [17]: which conducts SVD on sample by term matrix first and starting from most significant left eigenvector, and select samples with highest entry value.
- NMF [19]: which performs NMF on sample by term matrix and select samples best represent the discovered bases.
- SNMF [22]: which constructs the sample-sample similarity matrix first, clusters all samples with Symmetric Non-negative Matrix Factorization (SNMF) and extracts centering sentences from the clusters.
- KMEANS [13]: which performs K-means clustering over the dataset, and samples nearest to cluster centers are selected.
- NCUT [33]: which is similar to KMEANS, while use normalized cut as clustering method.

Besides, the following text-based microblog summarization approaches are also compared.

- PR [34]: the Phrase Reinforcement algorithm, which generates summaries by looking for the most commonly occurring phrases.
- HTF-IDF [6]: which selects summary posts by their Hybrid TF-IDF weights, and filters redundant posts with similarity threshold.
- CLUSTER [6]: another method proposed by Inouye *et al.* [6]. Similar to the traditional clustering-based multi-document summarization approach, this method first conducts kmeans $++$ to cluster the data samples. When selecting summary posts from each cluster, the above HTF-IDF is utilized to assign weights to the samples.

For our proposed approach, two specific methods are evaluated for comparison:

- MMES: the proposed multimedia social event summarization (MMES) framework that uses both text and visual contents in building CMLDA model.
- MMES-I: MMES without utilizing the visual information. In the subevent discovery stage, when applying CMLDA model, all microblog samples are assumed to be comprised of texts only.
- MMES-R: MMES without the process of irrelevant image removal, where the whole noisy image collections are used. This is the method adopted in our previous work [11].

TABLE II
COMPARISON AMONG DIFFERENT SUMMARIZATION APPROACHES ON THE *SOCIAL TRENDS* DATASET. AVERAGE RESULTS OF THE 20 EVENTS ARE REPORTED FOR ALL EVALUATION MEASUREMENT

System	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-SU
RANDOM	0.2453	0.0759	0.0217	0.0693
LSA	0.4039	0.1562	0.0367	0.1400
NMF	0.3119	0.1010	0.0251	0.0860
SNMF	0.3243	0.1598	0.0294	0.1377
KMEANS	0.3279	0.0985	0.0285	0.0951
NCUT	0.3761	0.1385	0.0357	0.1336
PR	0.3333	0.1564	0.0287	0.1366
HTF-IDF	0.3478	0.1306	0.0347	0.1558
CLUSTER	0.3586	0.1807	0.0357	0.1225
MMES-I	**0.4503**	**0.2419**	**0.0529**	**0.1761**
MMES-R	**0.4793**	**0.2631**	**0.0589**	**0.1997**
MMES	**0.5049**	**0.3076**	**0.0696**	**0.2356**

TABLE III
COMPARISON AMONG DIFFERENT SUMMARIZATION APPROACHES ON THE *PRODUCT EVENTS* DATASET. AVERAGE RESULTS OF THE 20 EVENTS ARE REPORTED FOR ALL EVALUATION MEASUREMENT

System	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-SU
RANDOM	0.2534	0.1210	0.0734	0.0741
LSA	0.3681	0.1886	0.1084	0.1265
NMF	0.3105	0.1546	0.0882	0.0923
SNMF	0.3079	0.1507	0.0888	0.0945
KMEANS	0.3336	0.1750	0.0977	0.0987
NCUT	0.3691	0.1747	0.1044	0.1271
PR	0.3150	0.1598	0.0917	0.0950
HTF-IDF	0.3215	0.1456	0.0956	0.1056
CLUSTER	0.3565	0.1621	0.1002	0.1121
MMES-I	**0.4223**	**0.2271**	**0.1196**	**0.1653**
MMES-R	**0.4533**	**0.2421**	**0.1256**	**0.1751**
MMES	**0.4780**	**0.2797**	**0.1492**	**0.1877**

The overall comparison of proposed MMES, MMES-I and MMES-R with other approaches are shown in Table II and Table III. In addition, detailed ROUGE-1 performance for each event is shown in Fig. 6. For conciseness, only seven selected comparing methods are shown in the figure. As can be seen from the results, the proposed MMES outperforms other methods for all events as well as all evaluation measurements. The good performance of MMES benefits from the following three aspects.

First of all, MMES explores the joint correlation between the textual and visual aspects of microblogs. The impact of multimedia knowledge can be demonstrated by comparing the results of MMES and MMES-I. The latter approach differs from MMES only with the lack of visual component. The performance illustrates the degradation of summarization ability when only a single media type is used. In addition, by comparing the results of MMES and MMES-R (which uses the noisy image collections), it clearly demonstrates the necessity for removing irrelevant images from the original datasets.

Secondly, MMES and MMES-I discover subevents before the summarization procedure. As a result, all important branches for an event are covered in the final summarization. Although some comparing methods also consider the coverage of the summarization for the dataset, the coverage is only considered at the event-level rather than the subevent-level. In case a subevent contains a small number of microblogs, there is a high probability that the microblogs related to this subevent will be ignored with comparing methods. The high performance of MMES-I as compared to all the baseline methods demonstrates the effectiveness of subevent discovery for enhancing summarization performance.

Thirdly, three criteria are specified in MMES for generating the summary of each subevent, namely coverage, significance and diversity. These three criteria are able to further facilitate the summary generation. We conduct further experiment to evaluate the effectiveness of each individual component by removing each of the three criteria from our framework. The result is shown in Table IV and Table V. MMES-C denotes the method of using only significance and diversity, without taking coverage into consideration. Similarly, MMES-S is the method without considering significance, and MMES-D represents our method without considering diversity. For each comparing methods, the parameter ω corresponding to the removed criterion was set to 0, while the parameters for other two factors were kept unchanged (The parameter value is described in the next subsection). As can be seen, the performance of removing any criterion becomes worse, which illustrates that all components are necessary for our framework.

An example of our summarization result is shown in Fig. 9. This is a summary on Event #1 of *Social Trends* dataset. As shown, five subevents are discovered. Due to space limitation, only the top 3 images and top 5 texts for each subevent are listed. This example demonstrates the ability of our proposed framework in: 1) well organizing the messy microblogs into structured subevents; 2) generating high-quality textual summary at subevent level; and 3) selecting the most representative images for summarizing the event.

D. Parameter Tuning

The overall selection score is a weighted linear combination of the three criteria coverage, significance and diversity. In this part, we examine the effects of the corresponding weighting parameters ω_1, ω_2 and ω_3 to achieve the optimal parameter setting. Keeping other two parameters fixed to 1, we vary the remaining ω from 0 to 10 to examine its influence on the final results, and select the value which achieves the best F-score for the ROUGE values. After achieving the corresponding values for ω_1, ω_2 and ω_3, we adjust $\omega_i = \omega_i/(\omega_1 + \omega_2 + \omega_3)$ to make the sum of the three weighting parameters to 1. With this procedure, the parameters are selected as $\omega_1 = 0.4$, $\omega_2 = 0.4$ and $\omega_3 = 0.2$ for the *Social Trends* dataset, and $\omega_1 = 0.2$, $\omega_2 = 0.5$, and $\omega_3 = 0.3$ for the *Product Events* dataset. In order to prove the above results are the optimized combination, we further fix two of the ω values fixed as the achieved value, and vary the third one. According to the results shown in Fig. 7, all parameters perform the best when they are at the achieve optimized value, e.g., the best performance for ω_1 in *Social Trends* dataset is 0.4, which

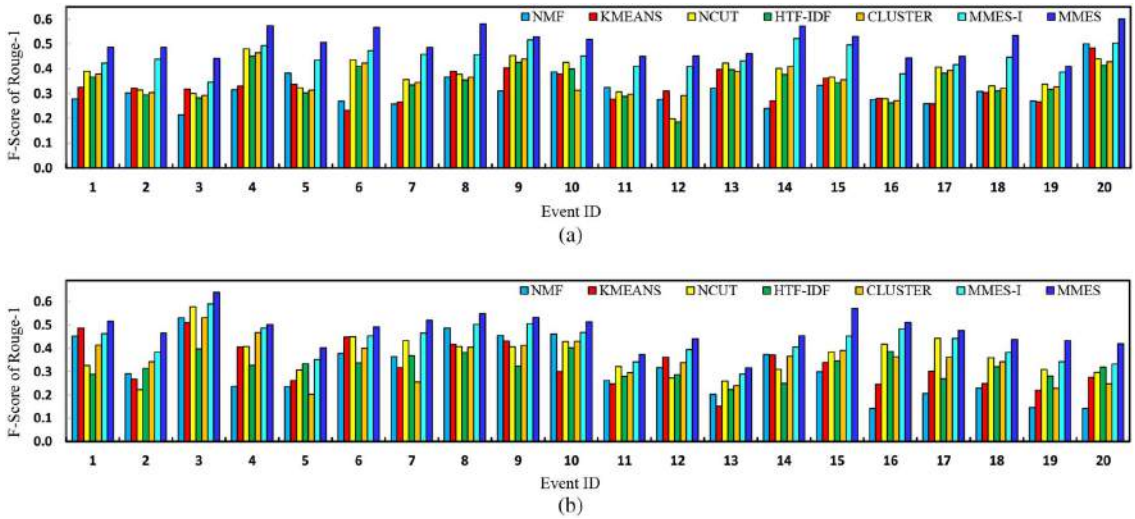


Fig. 6. Detailed performance (ROUGE-1) of MMES, MMES-I, and five selected comparing approaches over all events. (a) *Social Trends*. (b) *Product Events*.

TABLE IV
EFFECTS OF COVERAGE, SIGNIFICANCE AND DIVERSITY CRITERIA IN SUBEVENT DISCOVERY ON THE *SOCIAL TRENDS* DATASET

	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-SU
MMES-C	0.3602	0.2108	0.0335	0.1340
MMES-S	0.3791	0.2225	0.0327	0.1660
MMES-D	0.4207	0.2469	0.0322	0.1843
MMES	**0.5049**	**0.3076**	**0.0696**	**0.2356**

TABLE V
EFFECTS OF COVERAGE, SIGNIFICANCE AND DIVERSITY CRITERIA IN SUBEVENT DISCOVERY ON THE *PRODUCT EVENTS* DATASET

	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-SU
MMES-C	0.3502	0.2034	0.0506	0.1023
MMES-S	0.3701	0.1872	0.0823	0.1205
MMES-D	0.4213	0.2356	0.0956	0.1203
MMES	**0.4780**	**0.2797**	**0.1492**	**0.1877**

is consistent with our result, thus proves the optimization of the tuned parameter values.

Another important parameter is the number of subevents K. Fig. 8 shows the performance of MMES with various subevent number K in terms of ROUGE-1 result. Very small K fails to achieve satisfactory performance, as the ability to discover subevents is not fully utilized in this situation. However, large K does not lead to significant growth for the summarization performance, and may exert negative influence. By taking a detailed observation of our dataset, we can see that the microblog discussion for the same event is usually limited to a few directions, which means the number of subevents will not be too large in our specific scenario. If we set the subevent number to an improper large number, less important topic branches will be extracted, and corresponding microblogs will be included in the final summary, which will hurt the summarization performance. Furthermore, too many subevents will hurt the "concise" principle of summarization. Taken the above points into consideration, we set the subevent number K to 10 for *Social Trends* and 7 for *Product Events*.

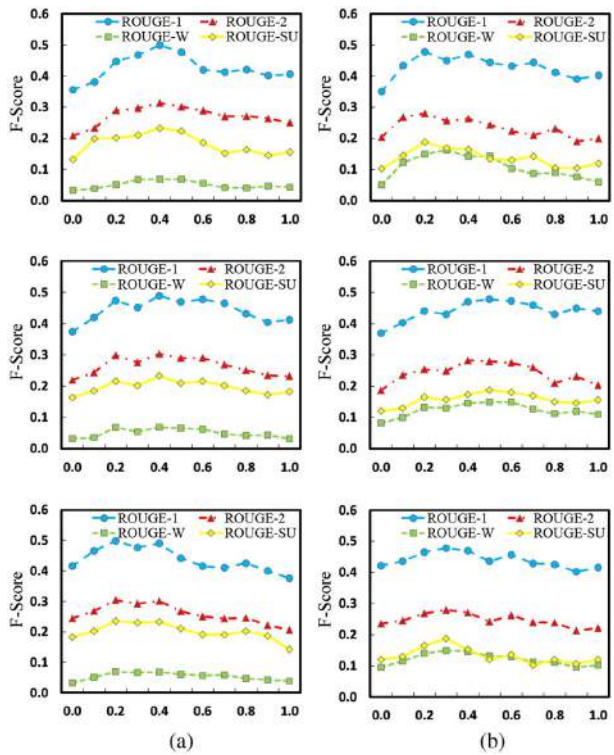


Fig. 7. Performance of parameter $\omega_1\omega_2$ and ω_3 on the two datasets. (a) *Social Trends*. (b) *Product Events*.

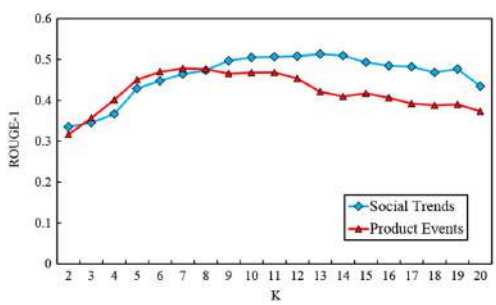


Fig. 8. Summarization performance of MMES with various subevent number K.

Fig. 9. Illustrative example of multimedia social event summarization on Event #1 in *Social Trends* dataset.

VI. Conclusion

In this paper, we present a multimedia social event summarization framework which automatically generates holistic visualized summary from the microblogs of various media types. The proposed framework features the exploration of the intrinsic correlations among different media types for enhancing the summarization performance. In particular, we developed three major stages to accomplish the summarization. First, we devise an effective approach for eliminating the potentially noisy images from raw microblog image collection. Then, we proposed a novel *Cross-Media-LDA* (CMLDA) model, to discover subevents from microblogs of different media types. Finally, we generated multimedia summary for social events utilizing the cross-media distribution knowledge of all the discovered subevents. We conducted extensive experiments on two real-world microblog datasets collected by ourselves to show the superiority of our proposed method as compared to the state-of-the-art approaches. In the future, we intend to extend the cross-media framework for automatically detecting social events and retrieving related candidate microblogs. In addition, we will also explore personalized microblog summarization based on user profile.

(a)

“第85届奥斯卡金像奖”，“Happy白色情人节”，“阳春三月，北京下雪”，“广药加多宝拼抢红罐归属权”，“柴静：《看见》”，“莫言荣获诺贝尔文学奖”，“美国春晚：2013超级碗”，“陕西房姐龚爱爱被爆有多处房产”，“聚美优品被指售假货”，“疯狂黄金周，丽江上万旅客打地铺”，“《西游降魔篇》上映”，“铁道部被取消”，“2013元宵晚会”，“好莱坞年度大片《云图》”，“祭奠逝去的生命小皓博”，“委内瑞拉总统查韦斯病逝”，“湖北荆州长江大桥卧铺车坠桥”，“雾霾天气，需戴口罩”，“埃及热气球坠毁9名港人遭意外”，“《笑傲江湖》大结局曝光”

(b)

“苹果WWDC开幕”，“微软正式发布WINDOWS 8预览版”，“微软OFFICE2013 PREVIEW发布会”，“诺基亚LUMIA发布会”，“墨西哥百事将推出新款MJ纪念罐”，“三星GALAXY 3 I9300 发布”，“HTC ONE 6月发布”，“深港澳车展”，“重庆车展”，“长春车展”，“迪奥魅惑香水在上海举行新品发布会”，“现代 朗动上市”，“雪铁龙DS5上市”，“法拉利F12 BERLI”，“克莱斯勒300C上市”，“本田CR-Z上市”，“马自达CX-5上市”，“奥迪Q3上市”，“本田艾力绅上市”，“汉兰达 2012”

Fig. 10. Contents of the events in the two datasets. (a) Event list of *Social Trends* dataset. (b) Event list of *Product Events* dataset.

APPENDIX

The contents of the events in the two datasets are listed in Fig. 10.

REFERENCES

[1] T.-S. Chua, H. Luan, M. Sun, and S. Yang, “Next: Nus-Tsinghua center for extreme search of user-generated content,” *IEEE MultiMedia Mag.*, vol. 19, no. 3, pp. 81–87, Jul.-Sep. 2012.

[2] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: Real-time event detection by social sensors,” in *Proc. WWW*, 2010, pp. 851–860.

[3] J. Weng and B.-S. Lee, “Event detection in Twitter,” in *Proc. ICWSM*, 2011.

[4] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua, “Emerging topic detection for organizations from microblogs,” in *Proc. SIGIR*, 2013, pp. 43–52.

[5] K. Spärck Jones, “Automatic summarising: The state of the art,” *Inf. Process. Manage.*, vol. 43, no. 6, pp. 1449–1481, 2007.

[6] D. Inouye and J. K. Kalita, “Comparing Twitter summarization algorithms for multiple post summaries,” in *Proc. SocialCom*, 2011, pp. 298–306.

[7] D. Chakrabarti and K. Punera, “Event summarization using tweets,” in *Proc. ICWSM*, 2011, pp. 66–73.

[8] B. Sharifi, M.-A. Hutton, and J. Kalita, “Summarizing microblogs automatically,” in *Proc. NAACL HLT*, 2010, pp. 685–688.

[9] Y. Yang, Z.-J. Zha, Y. Gao, X. Zhu, and T.-S. Chua, “Exploiting web images for semantic video indexing via robust sample-specific loss,” *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1677–1689, Oct. 2014.

[10] J. Bian, Y. Yang, and T.-S. Chua, “Predicting trending messages and diffusion participants in microblogging network,” in *Proc. SIGIR*, 2014, pp. 537–546.

[11] J. Bian, Y. Yang, and T.-S. Chua, “Multimedia summarization for trending topics in microblogs,” in *Proc. CIKM*, 2013, pp. 1807–1812.

[12] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, “Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion,” *Inf. Process. Manage.*, vol. 43, no. 6, pp. 1606–1618, 2007.

[13] D. R. Radev, H. Jing, M. Styś, and D. Tam, “Centroid-based summarization of multiple documents,” *Inf. Process. Manage.*, vol. 40, no. 6, pp. 919–938, 2004.

[14] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, “Summarizing text documents: Sentence selection and evaluation metrics,” in *Proc. SIGIR*, 1999, pp. 121–128.

[15] R. Mihalcea and P. Tarau, “A language independent algorithm for single and multiple document summarization,” in *Proc. IJCNLP*, 2005.

[16] G. Erkan and D. R. Radev, “Lexpagerank: Prestige in multi-document text summarization,” in *Proc. EMNLP*, 2004, pp. 365–371.

[17] Y. Gong and X. Liu, “Generic text summarization using relevance measure and latent semantic analysis,” in *Proc. SIGIR*, 2001, pp. 19–25.

[18] A. Haghighi and L. Vanderwende, “Exploring content models for multi-document summarization,” in *Proc. NAACL HLT*, 2009, pp. 362–370.

[19] S. Park, J.-H. Lee, D.-H. Kim, and C.-M. Ahn, “Multi-document summarization based on cluster using non-negative matrix factorization,” in *Proc. SOFSEM*, 2007, pp. 761–770.

[20] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, “Document summarization using conditional random fields,” in *Proc. ACL*, 2007, vol. 7, pp. 2862–2867.

[21] J. M. Conroy and D. P. O’Leary, “Text summarization via hidden Markov models,” in *Proc. SIGIR*, 2001, pp. 406–407.

[22] D. Wang, T. Li, S. Zhu, and C. Ding, “Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization,” in *Proc. SIGIR*, 2008, pp. 307–314.

[23] C. Lin, C. Lin, J. Li, D. Wang, Y. Chen, and T. Li, “Generating event storylines from microblogs,” in *Proc. CIKM*, 2012, pp. 175–184.

[24] Y. Yang, Y. Yang, and H. Shen, “Effective transfer tagging from image to video,” *TOMCCAP*, vol. 9, no. 2, 2013.

[25] Y. Yang, Y. Yang, Z. Huang, H. Shen, and F. Nie, “Tag localization with spatial correlations and joint group sparsity,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 881–888.

[26] R. Yan, X. Wan, M. Lapata, W. X. Zhao, P.-J. Cheng, and X. Li, “Visualizing timelines: Evolutionary summarization via iterative reinforcement between text and image streams,” in *Proc. CIKM*, 2012, pp. 275–284.

[27] P. Li, J. Ma, and S. Gao, “Learning to summarize web image and text mutually,” in *Proc. ICMR*, 2012, p. 28.

[28] W. Liu, Y.-G. Jiang, J. Luo, and S.-F. Chang, “Noise resistant graph ranking for improved web image search,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 849–856.

[29] T. Shi, M. Belkin, and B. Yu, “Data spectroscopy: Eigenspaces of convolution operators and clustering,” *Ann. Statist.*, vol. 37, no. 6B, pp. 3960–3984, 2009.

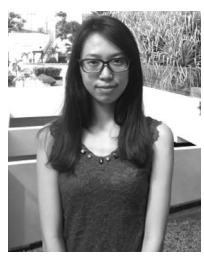
[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” in *Proc. JMLR*, 2003, vol. 3, pp. 993–1022.

[31] Y. Gao, F. Wang, H. Luan, and T.-S. Chua, “Brand data gathering from live social media streams,” in *Proc. ICMR*, 2014, p. 169.

[32] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out: ACL-04 Workshop*, 2004, pp. 74–81.

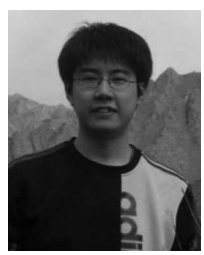
[33] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[34] B. Sharifi, M.-A. Hutton, and J. K. Kalita, “Experiments in microblog summarization,” in *Proc. SocialCom*, 2010, pp. 49–56.



Jingwen Bian received the B.S. degree in computer science from Peking University, Beijing, China, in 2010, and is currently working toward the Ph.D. degree at the School of Computing, National University of Singapore, Singapore.

Her research interest includes social network analysis and social media processing.



Yang Yang received the B.S. degree from Jilin University, Changchun, China in 2006, the M.E. degree from Peking University, Beijing, China, in 2009, and the Ph.D. degree from The University of Queensland, Brisbane, Qld., Australia, in 2013.

He was a Research Fellow with the School of Computing, National University of Singapore. He is currently with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include multimedia information retrieval, social media analysis, and machine learning.

Hanwang Zhang received the B.Eng. (Hons.) degree in computer science from Zhejiang University, Hangzhou, China, in 2009, and the Ph.D. degree in computer science from the National University of Singapore, Singapore, in 2014.

His main research interests include multimedia and computer vision, focusing on developing techniques for efficient search and recognition in image contents.

Tat-Seng Chua is currently the KITHCT Chair Professor with the School of Computing, National University of Singapore (NUS), Singapore, where he was the Acting and Founding Dean of the School of Computing from 1998 to 2000. He spent three years as a Research Staff Member with the Institute of Systems Science starting in 1980. He joined NUS in 1983. He is the Independent Director of two listed companies in Singapore. His main research interests include multimedia information retrieval, multimedia question answering, and the analysis and structuring of user-generated contents.

Dr. Chua has organized and served as a Program Committee Member of numerous international conferences in the areas of computer graphics, multimedia, and text processing. He was the Conference Co-Chair of ACM Multimedia in 2005, the Conference on Image and Video Retrieval in 2005, and the ACM SIGIR in 2008, and was the Technical PC Co-Chair of SIGIR in 2010. He serves on the Editorial Boards of the *ACM Transactions of Information Systems*, *Foundation and Trends in Information Retrieval*, *The Visual Computer*, and *Multimedia Tools and Applications*. He is on the Steering Committee of the International Conference on Multimedia Retrieval, Computer Graphics International, and Multimedia Modeling Conference Series. He serves as a Member of the international review panels of two large-scale research projects in Europe.