Cross-Platform Emerging Topic Detection and Elaboration from Multimedia Streams

BING-KUN BAO, CHANGSHENG XU, and WEIQING MIN, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
MOHAMMOD SHAMIM HOSSAIN, College of Computer and Information Sciences, King Saud University

With the explosive growth of online media platforms in recent years, it becomes more and more attractive to provide users a solution of emerging topic detection and elaboration. And this posts a real challenge to both industrial and academic researchers because of the overwhelming information available in multiple modalities and with large outlier noises. This article provides a method on emerging topic detection and elaboration using multimedia streams cross different online platforms. Specifically, Twitter, New York Times and Flickr are selected for the work to represent the microblog, news portal and imaging sharing platforms. The emerging keywords of Twitter are firstly extracted using aging theory. Then, to overcome the nature of short length message in microblog, Robust Cross-Platform Multimedia Co-Clustering (RCPMM-CC) is proposed to detect emerging topics with three novelties: 1) The data from different media platforms are in multimodalities; 2) The coclustering is processed based on a pairwise correlated structure, in which the involved three media platforms are pairwise dependent; 3) The noninformative samples are automatically pruned away at the same time of coclustering. In the last step of cross-platform elaboration, we enrich each emerging topic with the samples from New York Times and Flickr by computing the implicit links between social topics and samples from selected news and Flickr image clusters, which are obtained by RCPMM-CC. Qualitative and quantitative evaluation results demonstrate the effectiveness of our method.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

General Terms: Algorithms

Additional Key Words and Phrases: Topic detection, cross-platform, cross-media, coclustering

ACM Reference Format:
Bing-Kun Bao, Changsheng Xu, Weiqing Min, and Mohammod Shamim Hossain. 2015. Cross-platform emerging topic detection and elaboration from multimedia streams. ACM Trans. Multimedia Comput. Commun. Appl. 11, 4, Article 54 (April 2015), 21 pages.
DOI: http://dx.doi.org/10.1145/2730889

54

1. INTRODUCTION

Online media platforms are playing more and more important roles in exchanging and sharing emerging social topics in people's lives with the widespread usage and fast reaching speed [Aiello et al. 2013, 2015]. It was reported that Bolt inspired 80,000

This work was supported in part by National Program on Key Basic Research Project 973 Program, Project 2012CB316304, in part by the National Natural Science Foundation of China under Grant 61201374, Grant 61432019, and Grant 61225009, in part by Beijing Natural Science Foundation (4152053 and 4131004). The authors extend their appreciation to the Deanship of Scientific Research at King Saud University, Riyadh, Saudi Arabia for funding this work through the research group Project No. RGP VPP-228.
Authors' addresses: B.-K. Bao, C. Xu, and W. Min, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences; email: {bkbao,wqmin,csxu}@mail.nlpr.ia.ac.cn; M. S. Hossain, SWE Dept., College of Computer and Information Sciences, King Saud University; email: mshossain@ksu.edu.sa.

DOI: http://dx.doi.org/10.1145/2730889

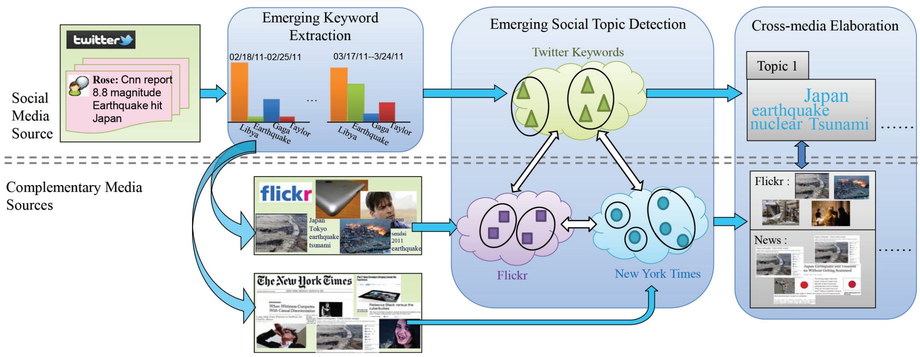


Fig. 1. The overview of emerging social topic detection and cross-platform elaboration.

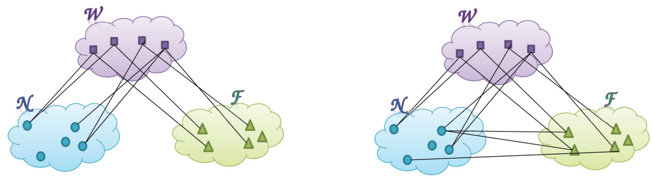
tweets per minute during his 200-meter race in London Olympics.[1] Therefore, how to dig up emerging social topics and elaborate them with vivid visual and textual contents all using those online media platforms becomes a hotter and hotter research area and receives attentions from both industrial and academic circles.

There are three major online media platforms in exchanging and sharing emerging social topics: microblogs (e.g., Twitter), news portals (e.g., *New York Times*), and image sharing websites (e.g., Flickr). Each of them has unique characteristics and advantages. Microblogs spread information fast and in real time, as the contents are user-generated. While they usually contain large mixed outlier noises and have a relatively short limit on message length. The news published on news portals are usually with better coverage, multiple-angel views and more creditable, but they are based on the opinion of an editorial board and not always reflect what are the real hot topics among social internet users at the moment. Image sharing websites like Flickr, though not really designed on spreading news, are hosting millions of photos, which are still rapidly increasing every second and could elaborate the hot topics circulating on micro-blogs and new portals. In this article, we exploit the best of all these three platforms to deliver a robust method of emerging topic detection and elaboration. Specifically, Twitter, New York Times and Flickr are selected, as they are among the best in their peers.

To accomplish the goal, we need to overcome three distinct challenges. (1) The overwhelming data in all platforms are inevitably mixed with plenty of nonemerging data, such as daily charts and uninformative messages. This makes the step of emerging data extraction necessary. (2) Since the data are in multiple modalities and from three heterogeneous yet complementary media platforms, revealing the relationships among those data is much more critical and challenging. (3) Based on the extracted emerging data and data relationships, a robust and effective topic detection and elaboration method is expected.

The proposed framework includes three stages: emerging keyword extraction, emerging social topic detection, and cross-media elaboration, as shown in Figure 1. In emerging keyword extraction stage, aging theory [Chen et al. 2007] is utilized in Twitter to extract the emerging keywords for each time interval (Section 3). Then, in stage of emerging social detection, we propose a novel multimedia topic detection method, called Robust Cross-Platform Multimedia Co-Clustering (RCPMM-CC), to simultaneously cocluster the multimodality data from three media sets respectively (Section 4).

[1]Salvador Rodriguez, "Twitter: More than 150 million tweets were sent about Olympics", Los Angeles Time, August 13, 2012. http://articles.latimes.com/2012/aug/13/business/la-fi-tn-twitter-2012-olympics-20120813.story.



(a) Star Structure

(b) Pairwise Correlated Structure

Fig. 2. Two kinds of structures on media platforms $\mathcal{W}, \mathcal{N}, \mathcal{F}$ in co-clustering.

The last stage is to enrich the emerging social topics, which are obtained as the top emerging Twitter keyword clusters, with the most relevant data from Flickr and The *New York Times* (Section 5). In this step, the implicit links between topics and samples can be revealed by the by-production from RCPMM-CC.

This work is an extension of our two previous work on social event/topic detection: RHOCC [Bao et al. 2013] and GS-HOCC [Bao et al. 2012]. RHOCC was proposed for single platform, for instance, Flickr, and coclusters images and the associated metadata, like timestamp, location, etc., to discover social events. As the applications are different, that is, RCPMM-CC for multiplatform vs. RHOCC for single platform, the two methods utilize different coclustering structures. RHOCC is based on the star structure, in which the auxiliary domains are pairwise independent while each of them is correlated with the central domain, referred to Figure 2(a). RCPMM-CC is based on the pairwise correlated structure, in which all the domains involved are pairwise dependent, referred to in Figure 2(b). GS-HOCC is an initial attempt to solve the problems similar to this work. There are two main differences with the proposed work. On one hand, GS-HOCC does not touch on how to process the uninformative samples, which makes it not as robust as this work, and therefore limits outreach to real life applications. On the other hand, GS-HOCC is based on information-theoretic coclustering, while RPCMM-CC is based on Bregman information coclustering, which is more general and easy to apply in other applications with different Bregman divergences and coclustering bases, such as microarray data analysis squared Euclidean distance and $\mathcal{C}_6$ basis, and transfer learning applications with I-divergence or square Euclidean distance and $\mathcal{C}_2$ basis.[2]

The rest of this article is organized as follows. Section 2 reviews related work. Sections 3, 4, and 5 respectively present three steps of the proposed framework, that is, Emerging Keyword Extraction, Emerging Social Topic Detection, and Cross-Media Elaboration. Experimental results are reported in Section 6 and 7. We conclude the article with future work in Section 8.

2. RELATED WORK

As our work involves topic detection, topic elaboration, and coclustering, the review of related work is focused on these three areas.

2.1. Topic Detection

Detecting emerging topics on social or news streams has recently been a hot area to study for both industrial and academic communities. There are several workon detecting topics from single media platform. Digg,[3] a social news website, receives a huge success by dynamically refreshing the most popular news. But it heavily relies on the

[2]As this article only discusses topic detection but not the extension of the proposed method, the experiments are focused on measuring the performance of removing outlier noises during coclustering.
[3]http://digs.com.

users' contribution. Sakaki et al. [2010] constructed an earthquake reporting system, in which each Twitter user is regarded as a "sensor". This work was constrained to detect preassigned events by training a classifier for each event. Chen et al. [2003, 2007] applied aging theory to model a news topic's life span and suggested that a news topic can be considered as a life form that goes through a life cycle of birth, growth, decay and death, reflecting its popularity over time. Aiello et al. [2013] found that trending topic detection method based on n-grams cooccurrence and a time-dependent boost ($df - idf_t$) is much more fit to detect topics on broader scales. Kasiviswanathan et al. [2011] addressed the problem of identifying emerging topics through the use of dictionary learning on the news stream. Some other researchers focused on topic detection with multiple media platforms. For example, Alvanaki et al. [2011] developed an emerging topic detection system, called En Blogue, by identifying shifts in the correlations between taxonomic tags. Sayyadi et al. [2009] proposed a method for event detection using the cooccurrence of keywords from different media platforms. Osborne et al. [2012] employed Wikipedia to improve the quality of discovered events in Twitter. The idea of this work is close to ours, but the approach and the introduced complementary platforms are quite different: unlike clustering on Twitter and tracking the page views of Wikipedia, we cocluster on Twitter and two complementary media platforms, New York Times and Flickr. However, all above methods not only neglect the multimodalities in media streams, but also fail to consider the nature of short length message on microblogs, which brings difficulties in constructing the accurate similarities between the extracted keywords. Taking the Japan earthquake as an example, nearly all tweets, such as "Terrible, it's earthquake.," "Now, it's shaking.," include only one or two emerging keywords, which is much less than the number of keywords in typical news. Our method follows the spirit of emerging keyword extraction with aging theory, but the distinctive feature from all previous work lies in the consideration of short length nature of microblog type social streams and the multi-modality data from media platforms.

2.2. Topic Elaboration

There have been numerous efforts indeed for providing vivid and attractive content from other media platforms to elaborate the topics in a certain media platform. Some work studied on discovering the implicit links among the individual samples from multiple media platforms. Takama et al. [2006] used the difference between publication times of news articles and blog posts to decide on the existence of a link. Gamon et al. [2008] used a graph-based approach to create context for news articles out of blog posts. Bansal et al. [2012] designed a framework to identify the event of social interests, and the contents of them were enhanced by images and news only from the associated links of corresponding tweets. Later, some methods were proposed to enrich the knowledge extracted from one source platform with the samples from only one complementary platform. Tan et al. [2011] enabled browsing systems to search results with rich media information through hyperliking of visual and textual snippets mined from multiple online resources. Roy et al. [2012] proposed a real-time transfer learning method to build an intermediate topic space in between social and video domains. Recently, some other work focused on topic enrichment with multimedia data from multiple media platforms to obtain comprehensive information, but there is still a long way to go. Serra et al. [2013] designed a framework, called STAMAT, to analyze topics in social networks, based both on textual and multimedia content. The topic in this work was extracted from a provided item from social streams, and obviously this kind of topic might not be an emerging one. In Bao et al. [2012], a news detection and pushing system was proposed to enrich the detected social events with news and images. However, this system does not process the uninformative samples, and surely is not robust in the topic detection step.

2.3. Coclustering

Coclustering has recently received a lot of attention as it is a good method to simultaneously cluster heterogeneous yet correlated metadata. Dhillon [2001] proposed a spectral coclustering algorithm by finding minimum cut vertex partitions in a bipartite graph between two heterogeneous types of metadata. However, the limitation of this method is that each metadatum with a certain type needs to be associated with at least of one metadatum with another type. Matrix factorization based coclustering methods coclusters both words and documents by data reconstruction through nonnegative matrix factorization on word-document matrix [Xu et al. 2003; Long et al. 2005; Ding et al. 2006]. This kind of method was then extended to tensor factorization on sets-relationship tensor for coclustering on three or more sets [Jegelka et al. 2009; Sang et al. 2012]. The constructed set-relationship tensor seeks the co-occurrence of three set samples, however, in our problem, this kind of relationship neglects the pairwise correlations between any two media platforms. Dhillon et al. [2003] proposed information-theoretic coclustering approach by finding a pair of maps from rows to rows clusters and from columns to column clusters with minimum mutual information loss on two heterogeneous types of metadata [Bao et al. 2015]. Later, Banerjee et al. [2007] suggested a generalization maximum entropy coclustering approach by appealing to Bregman information principle. To handle the noisy data in real world datasets, Deodhar et al. [2008] proposed Robust Overlapping Co-Clustering (ROCC) on two types of metadata, which provides clue to our work. To coclustering on three or more types of metadata, Gao et al. [2006] and Greco et al. [2010] dedicated on multiple types of heterogeneous data using high-order coclustering on star structure, in which there is a central metadata that connects any other types metadata to form a star structure of the interrelationship. Bao et al. [2013] proposed Robust High-Order Co-Clustering (RHOCC) method, which is also based on star structure, on Flickr images and their associated four types of modalities for social event detection.

Our work has three characteristics: (1) cross-platform processing ability with multimodality data; (2) applicability to real-world application via a novel pairwise correlated coclustering structure; and (3) robustness to outlier samples. To the best of our knowledge, there is no existing method on dealing with this kind of data.

3. EMERGING TWITTER KEYWORDS EXTRACTION

Table I shows the key notations in this article. For the emerging keyword extraction, we follow the proposed method of aging theory in Chen et al. [2007], that is, each keyword is considered as a live being, and the considered time period is divided into time intervals by a given time range r.

When a new keyword w appears, its life begins. At each following time interval, this keyword receives new nutrition according to its popularity as follows [Chen et al. 2007]:

$$newNutri_w(k) = |F_w(k)| \exp \frac{n_w(k)}{N(k)}, \qquad \text{where} \qquad |F_w(k)| = \frac{F_w(k)}{\sqrt{\sum_{k=1}^{|s|} F_w^2(k)}}, \tag{1}$$

$F_w(k)$ is the total frequency of the w-th keyword appearing in all the tweets published in the k-th time interval. $N(k)$ is the total number of tweets in the k-th time interval, while $n_w(k)$ indicates the number of tweets including w in the k-th time interval. Then, the new obtained nutrition converts into energy that can prolong the keyword's life.

Table I. List of Key Notations

Notation	Description
$\mathcal{W}, \mathcal{N}, \mathcal{F}$	collections of Twitter keywords, news and images
k, l, c	the numbers of informative samples in $\mathcal{W}$, $\mathcal{N}$ and $\mathcal{F}$
$\mathcal{W}_s = \{w_1, \ldots, w_k\}$, $\mathcal{N}_s = \{n_1, \ldots, n_l\}$ $\mathcal{F}_s = \{f_1, \ldots, f_c\}$	sets of informative samples' indexes in $\mathcal{W}$, $\mathcal{N}$ and $\mathcal{F}$
$\hat{u}, \hat{v}, \hat{h}$	the numbers of clusters of $\mathcal{W}_s$, $\mathcal{N}_s$ and $\mathcal{F}_s$
$\hat{\mathcal{W}} = \{\hat{w}_1, \ldots, \hat{w}_k\}$, $\hat{\mathcal{N}} = \{\hat{n}_1, \ldots, \hat{n}_l\}$ $\hat{\mathcal{F}} = \{\hat{f}_1, \ldots, \hat{f}_c\}$	sets of corresponding cluster's index for each sample in $\mathcal{W}_s$, $\mathcal{N}_s$ and $\mathcal{F}_s$
$W_s, N_s, F_s, \hat{W}, \hat{N}, \hat{F}$	random variables taking values in $\mathcal{W}_s, \mathcal{N}_s, \mathcal{F}_s, \hat{\mathcal{W}}, \hat{\mathcal{N}}, \hat{\mathcal{F}}$
$\mathbf{Z}, \mathbf{M}, \mathbf{Q}$	joint probability matrices between $\mathcal{W}$ and $\mathcal{N}$, $\mathcal{N}$ and $\mathcal{F}$, $\mathcal{W}$ and $\mathcal{F}$
$\hat{\mathbf{Z}}, \hat{\mathbf{M}}, \hat{\mathbf{Q}}$	joint probability matrices between $\hat{\mathcal{W}}$ and $\hat{\mathcal{N}}$, $\hat{\mathcal{N}}$ and $\hat{\mathcal{F}}$, $\hat{\mathcal{W}}$ and $\hat{\mathcal{F}}$

The conversion function is as follows:

$$newEng_w(k) = \frac{newNutri_w(k)}{1 + newNutri_w(k)}. \tag{2}$$

As the life being of keyword w grows, its energy will decay with age. Let $decNutri$ be the decay nutrition factor [Chen et al. 2003], which is an empirical constant. At the end of time interval, when $remEng_w(k-1) + newEng_w(k) > decNutr$, the remaining energy is

$$remEng_w(k) = remEng_w(k-1) + newEng_w(k) - d. \tag{3}$$

Otherwise, $remEng_w(k) = 0$. Finally, if the energy becomes not sufficient, the life of this keyword ends. With these predefined variables and functions, the whole procedure of emerging keyword extraction can be referred in Section 3 in Chen et al. [2007].

The energy variance of the keyword can be quantified how topical it is, refer to Equation (4):

$$a_w^{keyword}(k) = \sqrt{\frac{1}{|t|} \sum \left(remEng_w(k) - \overline{remEng}_w\right)^2}, \tag{4}$$

where $a_w^{keyword}(k)$ indicates the emerging weight of keyword w in the k-th time interval, $|t|$ is the number of time intervals, and $\overline{remEng}_w$ is the average remaining energy of keyword w. By sorting all the keywords' emerging weight in a certain time interval, we can get the corresponding top emerging keywords, which can be utilized to detect and sort the emerging topics.

4. CROSS-PLATFORM EMERGING TOPICS DETECTION

Due to the short length nature, the similarities between the detected keywords from Twitter are hardly to be revealed. Without accurate similarities, the clustering performance will be naturally unsatisfied. Thus, we introduce another two media platforms, *New York Times* and Flickr, as complementary to Twitter keywords, and simultaneously cocluster data from these three platforms respectively.

Obviously, our problem is different from the traditional coclustering problems in the following three aspects: (1) the data in our problem are multimodalities and from three disparate media platforms; (2) three media platforms are pairwise dependent because when a new emerging event occurs, the related information will be uploaded to these three media platforms; (3) only part of the data in three media platforms would form the cohesive clusters because they are mixed with a large amount of non-informative data. In this section, we will first give the definition of our problem, then respectively introduce our solutions to the above mentioned three issues in Section 4.2, Section 4.3, and Section 4.4 respectively. At last, we will present our proposed RCPMM-CC algorithm for emerging topic detection.

4.1. Problem Definition

The target of our work is to simultaneously discover the clusters of multimedia data from three pairwise dependent media platforms, and meanwhile prune away non-informative samples. Specifically, let $\mathcal{W}$, $\mathcal{N}$ and $\mathcal{F}$ be the set of data from extracted emerging Twitter keywords, New York Times and Flickr, and $\mathcal{W}_s$, $\mathcal{N}_s$ and $\mathcal{F}_s$ be the index sets of informative samples in $\mathcal{W}$, $\mathcal{N}$ and $\mathcal{F}$. The goal of emerging topic detection is to find the $\mathcal{W}_s$, $\mathcal{N}_s$, $\mathcal{F}_s$ from $\mathcal{W}$, $\mathcal{N}$, $\mathcal{F}$, and meanwhile map $\mathcal{W}_s$, $\mathcal{N}_s$, $\mathcal{F}_s$ into $\{1, \ldots, \hat{u}\}, \{1, \ldots, \hat{v}\}, \{1, \ldots, \hat{h}\}$ respectively, where $\hat{u}, \hat{v}, \hat{h}$ indicate the cluster numbers of $\mathcal{W}_s$, $\mathcal{N}_s$ and $\mathcal{F}_s$. We define $\hat{\mathcal{W}}$, $\hat{\mathcal{N}}$, $\hat{\mathcal{F}}$ as the sets of corresponding cluster's index for each sample in $\mathcal{W}_s$, $\mathcal{N}_s$, $\mathcal{F}_s$.

4.2. Joint Probabilities for Multi—Modality data on Different Platforms

Considering that the extracted emerging Twitter keywords can reveal the trends of social users' interests, our work only focuses on this emerging Twitter keyword set, and discards the images or videos that are associated to the original tweets. For data from New York Times and Flickr, it is obvious that there are two kinds of media modalities associated, that is, text and image. The image modality in New York Times is referred to the attached images, while text modality in Flickr is titles, surrounding texts and tags. Both textural and visual contents can contribute to discovering the correlations between the samples from different media platforms. Let $\mathbf{x}_w = (\mathbf{x}^t_w) \in \mathcal{W}$, $\mathbf{x}_f = (\mathbf{x}^t_f, \mathbf{x}^i_f) \in \mathcal{F}$ and $\mathbf{x}_n = (\mathbf{x}^t_n, \mathbf{x}^i_n) \in \mathcal{N}$, where $\mathbf{x}^t_w$, $\mathbf{x}^t_n$ and $\mathbf{x}^t_f$ indicate the text features of samples from Twitter keywords, New York Times and Flickr, while $\mathbf{x}^i_n$ and $\mathbf{x}^i_f$ indicate the corresponding visual features of samples from New York Times and Flickr. $\mathbf{x}^i_n = \mathbf{0}$ means no images attached in news sample.

As the correlations among the data from different platforms are the fundamentals for coclustering, we utilize the discrete joint probabilities on each of the two platforms to measure these correlations. Considering two modalities of data from New York Times and Flickr, the correlations between them should fuse the textual similarity and visual similarity. Let $\mathbf{S}^t_{WF}$, $\mathbf{S}^t_{WN}$, $\mathbf{S}^t_{NF}$ be textual similarities, and $\mathbf{S}^i_{WF}$, $\mathbf{S}^i_{WN}$, $\mathbf{S}^i_{NF}$ be the visual ones.

$$\mathbf{S}^t_{WF}(w, f) = e^{-[\mathbf{x}^t_w(\mathbf{x}^t_f)^T/(\sqrt{2}\theta\|\mathbf{x}^t_w\|\|\mathbf{x}^t_f\|)]^2}, \quad \mathbf{S}^t_{WN}(w, n) = e^{-[\mathbf{x}^t_w(\mathbf{x}^t_n)^T/(\sqrt{2}\theta\|\mathbf{x}^t_w\|\|\mathbf{x}^t_n\|)]^2},$$
$$\mathbf{S}^t_{NF}(n, f) = e^{-[\mathbf{x}^t_n(\mathbf{x}^t_f)^T/(\sqrt{2}\theta\|\mathbf{x}^t_n\|\|\mathbf{x}^t_f\|)]^2}, \quad \mathbf{S}^i_{WF}(w, f) = 0, \quad \mathbf{S}^i_{WN}(w, n) = 0, \tag{5}$$

$$\mathbf{S}^i_{NF}(n, f) = \begin{cases} e^{-[\|\mathbf{x}^i_n - \mathbf{x}^i_f\|/(\sqrt{2}\theta)]^2} & \mathbf{x}^i_n \neq \mathbf{0} \\ 0 & \mathbf{x}^i_n = \mathbf{0} \end{cases},$$

where θ is a scaling parameter and adaptively assigned as the median value of all pair-wise Euclidean distances between two features. $(\cdot)^T$ is the matrix's transpose, superscript t and i are referred to text and image modalities. $S^i_{WN} = 0, S^i_{WF} = 0$ since no image is attached in $\mathcal{W}$.

To get the correlation among three platforms, we need to fuse the similarities from image and text modalities. Define the weights τ^i and τ^t for balancing these two terms, and

$$\delta(\mathbf{a}, \mathbf{b}) = \begin{cases} 0 & \mathbf{a} = \mathbf{0} \; or \; \mathbf{b} = \mathbf{0} \\ 1 & otherwise, \end{cases} \tag{6}$$

then the correlations can be calculated by,

$$\mathbf{C}_{WF}(w,f) = \mathbf{S}^t_{WF}(w,f), \quad \mathbf{C}_{WN}(w,n) = \mathbf{S}^t_{WN}(w,n),$$

$$\mathbf{C}_{NF}(n,f) = \begin{cases} 0 & \delta(\mathbf{x}^i_n, \mathbf{x}^i_f) = 0 \; and \; \delta(\mathbf{x}^t_n, \mathbf{x}^t_f) = 0 \\ \frac{\tau^i \delta(\mathbf{x}^i_n, \mathbf{x}^i_f)\mathbf{S}^i_{NF}(n,f) + \tau^t \delta(\mathbf{x}^t_n, \mathbf{x}^t_f)\mathbf{S}^t_{NF}(n,f)}{\tau^i \delta(\mathbf{x}^i_n, \mathbf{x}^i_f) + \tau^t \delta(\mathbf{x}^t_n, \mathbf{x}^t_f)} & otherwise, \end{cases} \tag{7}$$

Then the probabilities between $\mathcal{W}$ and $\mathcal{N}$, $\mathcal{N}$ and $\mathcal{F}$, $\mathcal{W}$ and $\mathcal{F}$ are

$$\mathbf{Z} = \frac{\mathbf{C}_{WN}}{\sum_{w,n}\mathbf{C}_{WN}}, \quad \mathbf{M} = \frac{\mathbf{C}_{NF}}{\sum_{n,f}\mathbf{C}_{NF}}, \quad \mathbf{Q} = \frac{\mathbf{C}_{WF}}{\sum_{w,f}\mathbf{C}_{WF}}. \tag{8}$$

4.3. Coclustering on Pairwise Correlated Structure

Since our work is much related to Bregman coclustering [Banerjee et al. 2007], we first briefly review the basic ideas of it.

4.3.1. Bregman Coclustering. For simplification, we take two platforms, $\mathcal{W}$ and $\mathcal{N}$, as an example in this part. Let ϕ be a real-valued strictly convex function defined on the convex set $\mathcal{S} = dom(\phi) \in \Re^d$, such that ϕ is differentiable on $\mathcal{S}$. *Bregman divergence* is defined as

$$d_\phi(z_1, z_2) = \phi(z_1) - \phi(z_2) - < z_1 - z_2, \nabla\phi(z_2) >, \tag{9}$$

where $\nabla\phi$ is the gradient of ϕ.

Define $\mathbf{Z}$ as the joint probability matrix on $\mathcal{W}$ and $\mathcal{N}$. Let $\nu = \{\nu_{wn}, w = 1, \ldots, u, n = 1, \ldots, v\}$ be the joint probability measure of the random variable pair (W, N), which is either pre-specified or set to be the uniform distribution. Let Z be a random variable taking values in $\mathbf{Z}$ following ν, that is $p(Z(W,N) = z_{wn}) = \nu_{wn}$. The *Bregman information* of Z defined as the expected Bregman divergence with respect to ν, that is, $I_\phi(Z) = E_\nu[d_\phi(Z, E_\nu(Z))]$, where $E_\nu(\cdot)$ is the expectation with respect to ν.

Similarly, let matrix $\hat{\mathbf{Z}} = \{\hat{z}_{wn}\}$ be an approximation for the matrix $\mathbf{Z}$ such that $\hat{\mathbf{Z}}$ depends only upon coclustering (ρ, γ) and certain summary statistics derived from the coclustering.[4] $\hat{Z}$ is defined as the random variable on $\hat{\mathbf{Z}}$ following ν. Then the goal of coclustering is to find the pair of coclustering mappings ρ and γ such that the expected Bregman distortion between Z and $\hat{Z}$ is minimized, that is

$$\min_{\rho,\gamma} E_\nu[d_\phi(Z, \hat{Z})] = \min_{\rho,\gamma} \sum_w \sum_n \nu_{wn} d_\phi(z_{wn}, \hat{z}_{wn}). \tag{10}$$

Specially, if $\phi(z) = z^2$, Equation (10) is $E_\nu[d_\phi(Z, \hat{Z})] = E_\nu[(Z - \hat{Z})^2]$.

4.3.2. Bregman Coclustering on Pairwise Correlated Structure. Recall that our problem involves three media platforms, $\mathcal{W}$, $\mathcal{N}$ and $\mathcal{F}$. The probabilities between $\mathcal{W}$ and $\mathcal{N}$, $\mathcal{N}$ and $\mathcal{F}$, $\mathcal{W}$ and $\mathcal{F}$ are $\mathbf{Z}, \mathbf{M}, \mathbf{Q}$ respectively. Let the triple (ρ, γ, ζ) be the coclustering mapping from $\mathcal{W}, \mathcal{N}$ and $\mathcal{F}$ to $\hat{u}, \hat{v}, \hat{h}$ clusters respectively. Figure 3 shows the coclustering process on $\mathcal{W}, \mathcal{N}$ and $\mathcal{F}$. In this figure, $\mathcal{W}, \mathcal{N}$ and $\mathcal{F}$ are pairwise dependent. In this figure, Mapping ρ is the coclustering on $\mathcal{W}$, that is, $\rho : \{w_1, w_2, w_3, w_4\} \rightarrow \{\{w_1, w_2\}, \{w_3, w_4\}\}$. Thus, ρ is $\mathbf{Z}$'s row clustering and $\mathbf{Q}$'s column clustering. Similarly, γ is $\mathbf{Z}_s$'s column clustering and $\mathbf{M}_s$'s row clustering, ζ is $\mathbf{M}_s$'s column clustering and $\mathbf{Q}_s$'s row clustering.

To seek the optimal (ρ, γ, ζ), we divide the coclustering into three subproblems according to pairwise structure, that is, optimizing (ρ, γ) by coclustering on $\mathbf{Z}$, optimizing

[4]The summary statistics is presented in Banerjee et al. [2007]. It may be properties of the coclusters themselves, such as cocluster marginals, and/or some other important statistics of data, such as row and column marginals.

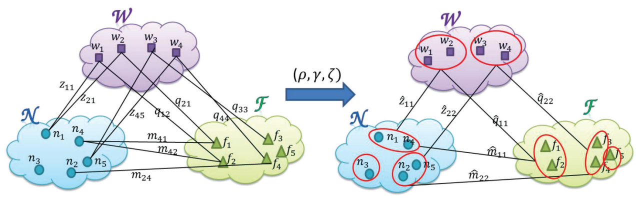


Fig. 3. The coclustering process of $\mathcal{W}, \mathcal{N}$ and $\mathcal{F}$.

(γ, ζ) by coclustering on $\mathbf{M}$ and optimizing (ζ, ρ) by coclustering on $\mathbf{Q}$. Obviously, if we conduct these three coclusterings independently, it will have a great probability that the clustering schemes for three media platforms are different in the different solutions. In other words, two results of ρ from coclusterings on $\mathbf{Z}$ and $\mathbf{Q}$ do not match in most cases. Therefore, we seek for three clustering mappings, such that each of them is not locally optimal, but the clustering results on three media platforms are the same, and the overall coclustering is globally optimal under a certain objective function. A simple but feasible objective function could just be the linear combination of three coclusterings.

Bregman coclustering provides us an effective way to cocluster on $\mathbf{Z}$, $\mathbf{M}$ and $\mathbf{Q}$ respectively. Let Z, M, Q be the random variables following ν, κ, ω[5], that is, $p(\hat{Z} = \hat{z}_{wn}) = \nu_{wn}$, $p(\hat{M} = \hat{m}_{nf}) = \kappa_{nf}$, and $p(\hat{Q} = \hat{q}_{fw}) = \omega_{wf}$. Let $\hat{\mathbf{Z}} = \hat{\mathbf{Z}}(\rho, \gamma)$, $\hat{\mathbf{M}} = \hat{\mathbf{M}}(\gamma, \zeta)$ and $\hat{\mathbf{Q}} = \hat{\mathbf{Q}}(\zeta, \rho)$ be the approximations to $\mathbf{Z}$, $\mathbf{M}$ and $\mathbf{Q}$ with a predefined approximation scheme, which is based on a coclustering basis.[6] $\hat{Z}$, $\hat{M}$, $\hat{Q}$ are defined as the random variables on $\hat{\mathbf{Z}}$, $\hat{\mathbf{M}}$, $\hat{\mathbf{Q}}$ following ν, κ, ω. Then the objective functions on three coclusterings are to minimize three expected Bregman distortions.

—Coclustering on $\mathbf{Z}$: $\min_{\rho,\gamma} E_\nu[d_\phi(Z, \hat{Z})]$
—Coclustering on $\mathbf{M}$: $\min_{\gamma,\zeta} E_\kappa[d_\phi(M, \hat{M})]$
—Coclustering on $\mathbf{Q}$: $\min_{\zeta,\rho} E_\omega[d_\phi(Q, \hat{Q})]$

Based on the above objective functions in three sub-problems, the overall can be formulated as the linear combination of these three expected Bregman distortions, as shown in Proposition 4.1.

PROPOSITION 4.1. *The quality of the Bregman coclustering on pairwise correlated structure is evaluated to minimize a linear combination of all the bi-clustering objective functions on correlated media platforms, that is,*

$$\alpha E_\nu[d_\phi(Z, \hat{Z})] + \beta E_\kappa[d_\phi(M, \hat{M})] + \xi E_\omega[d_\phi(Q, \hat{Q})] = \alpha \sum_w \sum_n \nu_{wn} d_\phi(z_{wn}, \hat{z}_{wn}) + \beta \sum_n \sum_f \kappa_{nf} d_\phi(m_{nf}, \hat{m}_{nf}) + \xi \sum_f \sum_w \omega_{fw} d_\phi(q_{fw}, \hat{q}_{fw}), \quad (11)$$

where $\alpha + \beta + \xi = 1$, and $d_\phi()$ denotes Bregman divergence.

4.4. Solution to Outlier Data

Robust Overlapping Co-Clustering (ROCC) proposed by Deodhar et al. [2008] extends Bregman coclustering to address the problem of large existence of noisy data. Inspired

[5]In our implementation, we set ν, κ, ω to be uniform distributions.
[6]There are six kinds coclustering bases, which are detailed in Banerjee et al. [2007]. In our implementation, we choose $\mathcal{C}_5$ as it achieves good performance in text clustering, which is also reported in Banerjee et al. [2007].

by this work, we discard the outlier data in $\mathcal{W}$, $\mathcal{N}$ and $\mathcal{F}$ by two steps. The first step is to discard the data with large approximation errors in minimizing the expected distortion between two random variables, one takes values in the original joint probability matrix, and the other takes values in the clustered joint probability matrix in Equation (10). In the second step, a post process is utilized to merge the similar clusters and prune away those with large errors.

Let k, l, and c be the specified numbers of $\mathcal{W}_s$, $\mathcal{N}_s$ and $\mathcal{F}_s$, that compose of the remaining data after pruning. The desired coclustering mapping pair (ρ, γ, ζ) is optimized by minimizing

$$\alpha \sum_{\hat{w}} \sum_{\hat{n}} \sum_{\rho(w)=\hat{w}} \sum_{\gamma(n)=\hat{n}} \nu_{wn} d_{\phi}(z_{wn}, \hat{z}_{wn}) + \beta \sum_{\hat{n}} \sum_{\hat{f}} \sum_{\gamma(n)=\hat{n}} \sum_{\zeta(f)=\hat{f}} \kappa_{nf} d_{\phi}(m_{nf}, \hat{m}_{nf}) + \xi \sum_{\hat{f}} \sum_{\hat{w}} \sum_{\zeta(f)=\hat{f}} \sum_{\rho(w)=\hat{w}} \omega_{fw} d_{\phi}(q_{fw}, \hat{q}_{fw}). \quad (12)$$

With this objective function, top k samples in $\mathcal{W}$, l samples in $\mathcal{N}$ and c samples in $\mathcal{F}$ with least errors are selected to form $\mathcal{W}_s$, $\mathcal{N}_s$ and $\mathcal{F}_s$.

Next, a postprocess is provided to adjust the obtained clusters by merging the similar ones and pruning those with large errors. A pair of coclusters, between which the distance is small, will be merged together into a new cocluster; while some coclusters with error values larger than a predefined "error cut-off value" will be filed out in this step (More specifically, please refer to Section 5.2 in Deodhar et al. [2008]).

4.5. Robust Cross-Platform Multimedia Coclustering Algorithm

The objective function Equation (12) can be optimized by an iterative procedure. If we fix γ and ζ, the best choice of ρ is to assign the w-th Twitter keyword to $\hat{w}^{\star}$ cluster by

$$\hat{w}^{\star} = \arg\min_{\hat{w}} \alpha E_{\nu}(\hat{w}) + \xi E_{\omega}(\hat{w}) = \arg\min_{\hat{w}} \alpha \sum_{\hat{n}} \sum_{\gamma(n)=\hat{n}} \nu_{wn} d_{\phi}(z_{wn}, \hat{z}_{wn}(\hat{w})) + \xi \sum_{\hat{f}} \sum_{\zeta(f)=\hat{f}} \omega_{fw} d_{\phi}(q_{fw}, \hat{q}_{fw}(\hat{w})). \quad (13)$$

After computing the best cluster assignment for each w, the top k keywords with minimum errors are remained to form the keyword clusters. The optimization of γ and ζ is similar to that of ρ.

The pseudocode for Robust Cross-Platform Multimedia Co-Clustering Algorithm (RCPMM-CC) is shown in Algorithm 1. The discussion of cluster number estimation computation complexity of RCPMM-CC is given in the Appendix, which is online via ACM Digital Library. From Algorithm 1, we can obtain the emerging keyword clusters, and the emerging weight for the k-th cluster can be calculated by

$$\hat{w}_k^{topic} = \frac{\sum_{i \in \mathcal{W}:\rho(i)=k} w_i^{keyword}}{|i|_{i \in \mathcal{W}:\rho(i)=k}}, \quad (14)$$

where $|i|_{i \in \mathcal{W}:\rho(i)=k}$ is the number of keywords in the k-th cluster.

5. CROSS-PLATFORM ELABORATION

Cross-platform elaboration aims to automatically enrich each detected emerging social topic with top relevant news and images from news portals and image sharing websites. With those supplemented contents, users are able to receive more comprehensive and credible information on the talk trends. In our implementation, the news portal and

ALGORITHM 1: Robust Cross-Platform Multimedia Co-Clustering Algorithm (RCPMM-CC):

Input:
$\mathbf{Z}, \mathbf{M}, \mathbf{Q}, k, l, c, \hat{u}, \hat{v}, \hat{h}$, basis $\mathcal{C}$, d_ϕ, α, β, ξ
Output:
$\hat{\mathcal{W}}, \hat{\mathcal{N}}, \hat{\mathcal{F}}$

Initialization: a random coclustering (ρ, γ, ζ)
Repeat
Update statistics for clusters $\hat{\mathcal{W}}, \hat{\mathcal{N}}, \hat{\mathcal{F}}$ based on basis $\mathcal{C}$ to compute new $\hat{\mathbf{Z}}, \hat{\mathbf{M}}, \hat{\mathbf{Q}}$;
For each w, find its new cluster as

$$\hat{w} = \arg\min_{\hat{w}} \alpha \sum_{\hat{n}} \sum_{\gamma(n)=\hat{n}} \nu_{wn} d_\phi(z_{wn}, \hat{z}_{wn}(\hat{w})) + \xi \sum_{\hat{f}} \sum_{\zeta(f)=\hat{f}} \omega_{fw} d_\phi(q_{fw}, \hat{q}_{fw}(\hat{w})). \tag{15}$$

$\hat{\mathcal{W}}$ = the set of k samples' indexes with least errors, and update ρ according to $\hat{\mathcal{W}}$.
For each n, find its new cluster as

$$\hat{n} = \arg\min_{\hat{n}} \alpha \sum_{\hat{w}} \sum_{\rho(w)=\hat{w}} \nu_{wn} d_\phi(z_{wn}, \hat{z}_{wn}(\hat{n})) + \beta \sum_{\hat{f}} \sum_{\zeta(f)=\hat{f}} \kappa_{nf} d_\phi(m_{nf}, \hat{m}_{nf}(\hat{n})). \tag{16}$$

$\hat{\mathcal{N}}$ = the set of l samples' indexes with least errors, and update γ according to $\hat{\mathcal{N}}$.
For each f, find its new cluster as

$$\hat{f} = \arg\min_{\hat{f}} \alpha \sum_{\hat{n}} \sum_{\gamma(n)=\hat{n}} \kappa_{nf} d_\phi(m_{nf}, \hat{m}_{nf}(\hat{f}) + \xi \sum_{\hat{w}} \sum_{\rho(w)=\hat{w}} \omega_{fw} d_\phi(q_{fw}, \hat{q}_{fw}(\hat{f})) \tag{17}$$

$\hat{\mathcal{F}}$ = the set of c samples' indexes with least errors, and update ζ according to $\hat{\mathcal{F}}$
until convergence
Prune coclusters with large errors
Merge similar coclusters until stopping criterion is reached
Return $\hat{\mathcal{W}}, \hat{\mathcal{N}}, \hat{\mathcal{F}}$.

image sharing websites are selected as New York Times and Flickr, which are the same as complementary platforms used in emerging social topic detection.

An intuitive approach for cross-platform elaboration is to directly link the datum from New York Times or Flickr with the detected emerging topic by different correlation weights, and the desired news and images are those with high weights. Note that the results from emerging social topic detection include not only the clusters of Twitter keywords, news and images, but also the joint probabilities on two clusters from different media platforms. Thus, we simplify the approach on linking a certain social topic with the data from high relevant news and images clusters instead of the whole news and image datasets to reduce the computational cost. Two steps are involved in our cross-platform elaboration: 1) identify the high relevant news and image clusters for each emerging social topic; 2) calculate the correlations between the datum from the selected clusters and the corresponding emerging social topic, and return the top ranked news and images.

For the i-th emerging topic, referred to $\hat{\mathcal{W}}_i \subset \mathcal{W}$, the joint probabilities of $\hat{\mathcal{W}}_i$ and all the New York Time clusters are denoted as $\hat{\mathcal{Z}}_i = \{\hat{z}_{i1}, \ldots, \hat{z}_{i\hat{v}}\}$, where $\hat{z}_{ij} = \sum_{w \in \mathcal{W}:\rho(w)=i} \sum_{n \in \mathcal{N}:\gamma(n)=j} z_{wn}$. We sort the elements in $\hat{\mathcal{Z}}_i$ by descending order, that is $\hat{\mathcal{Z}}_i = \{\hat{z}_{r_1}, \ldots, \hat{z}_{r_{\hat{v}}}\}, \hat{z}_{r_1} \geq \hat{z}_{r_2} \geq \cdots \geq \hat{z}_{r_{\hat{v}}}$. Then the top *cluNum* relevant New York Time clusters are remained, where *cluNum* is such that

$$\sum_{k=1}^{cluNum} \hat{z}_{r_k} \geq \epsilon \quad \text{and} \quad \sum_{k=1}^{cluNum-1} \hat{z}_{r_k} < \epsilon. \tag{18}$$

ϵ is a threshold value and set as 0.6 in our work. Define $\hat{\mathcal{N}}_i$ as the set of all the samples from the top *cluNum* relevant clusters. The correlation between this New York Times sample and the i-th emerging topic can be calculated as

$$\mathbf{C}_{\hat{W}N}(\hat{w}, n) = e^{-[\mathbf{x}^t_{\hat{w}_i}(\mathbf{x}^t_n)^T/(\sqrt{2}\theta\|\mathbf{x}^t_{\hat{w}_i}\|\|\mathbf{x}^t_n\|)]^2} \tag{19}$$

The top $\hbar_n$ news samples with the highest values in $\mathbf{C}_{\hat{W}N}(\hat{w}, \cdot)$ are finally selected as elaborated news samples to enrich the i-th emerging social topic.

For elaboration with Flickr image, the process to select top *cluNum* relevant Flickr clusters is similar as that in New York Times.

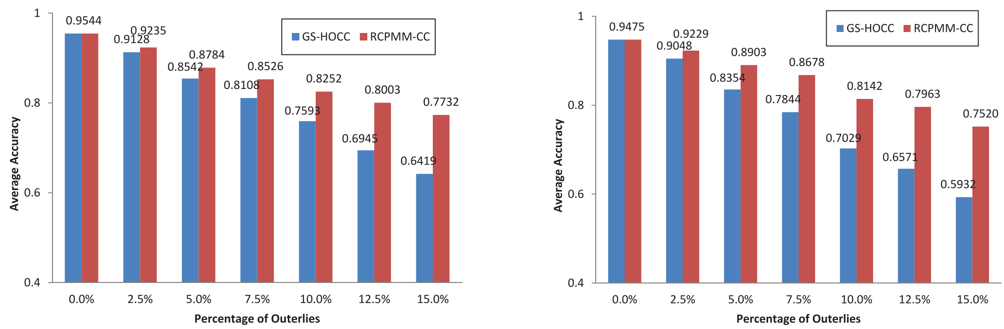
6. EXPERIMENTAL EVALUATION ON SYNTHETIC DATASETS

In order to measure the performance of proposed RCPMM-CC on processing dataset with outlier noise, we use synthetically generated datasets as the ground-truth co-cluster labels can be generated for the data matrix entries. The datasets are generated by the following procedure.

(1) Generate the numbers of clusters $\hat{u}$, $\hat{v}$ and $\hat{h}$. The values of $\hat{u}$, $\hat{v}$ and $\hat{h}$ are randomly selected from a uniform distribution (range 4–16).
(2) Generate joint probability matrices $\hat{\mathbf{Z}}, \hat{\mathbf{M}}$ and $\hat{\mathbf{Q}}$. Take $\hat{\mathbf{Z}} \in (0, 1)^{\hat{u}\times\hat{v}}$ $(\hat{u} \leq \hat{v})$ as an example, we first divide matrix $\hat{\mathbf{Z}}$ into two submatrices, one consisting of the first $\hat{u}$ columns, and the other consisting of last $\hat{u} - \hat{v}$ columns. For the first submatrix, the diagonal elements of it are randomly selected from a uniform distribution (range 0–1), and other elements are set to be 0. For the second submatrix, only two elements of each column are randomly set as nonzeros, and the selection of elements need to make the rank of this submatrix to be equal to $\hat{u} - \hat{v}$. At last, we normalize matrix $\hat{\mathbf{Z}}$ to make the sum of its elements as 1. Matrices $\hat{\mathbf{Z}}(\hat{u} > \hat{v})$ as well as $\hat{\mathbf{M}}$ and $\hat{\mathbf{Q}}$ are constructed in a similar way.
(3) Generate joint probability matrices $\mathbf{Z}, \mathbf{M}$ and $\mathbf{Q}$. First, we randomly select the element number of each cluster from a uniform distribution (range 20–40 for the first dataset and 80-100 for the second dataset). Second, we randomly split each element in $\hat{\mathbf{Z}}, \hat{\mathbf{M}}$ and $\hat{\mathbf{Q}}$ to a matrix whose dimension is equal to the corresponding cluster element number.
(4) Add outliers into $\mathbf{Z}, \mathbf{M}$ and $\mathbf{Q}$. We first set the number of noise samples added to $\mathcal{W}, \mathcal{N}$ and $\mathcal{F}$ as a given percentage of the number of informative samples, then sparsely choose the joint probability for the outliers and informative samples from uniform distribution (range 0–1), and at last, we normalize matrix $\mathbf{Z}, \mathbf{M}$ and $\mathbf{Q}$ to make the sum of their elements as 1 respectively.
(5) We permute the rows and columns of $\mathbf{Z}, \mathbf{M}$ and $\mathbf{Q}$ consistently.

Two synthetic datasets with different sizes are generated. The first one contains $162, 232$ and 287 informative samples in $\mathcal{W}$, $\mathcal{N}$ and $\mathcal{F}$ respectively, and those informative samples are from $5, 8$, and 10 clusters. The second one contains $661, 831, 1105$ informative samples from $7, 9, 12$ clusters respectively. The data size of the second one is around 4 times of the first one. For both datasets, we respectively added outlier samples with the size equal to percentages $(2.5\%, 5.0\%, 7.5\%, 10\%, 12.5\%, 15\%)$ of the informative ones.

We select GS-HOCC [Bao et al. 2012] as baseline because it is also based on information theoretic coclustering with pairwise correlated structure but does not pruning away outlier noises. The evaluation criteria is set as average accuracy over $\mathcal{W}$, $\mathcal{N}$ and $\mathcal{F}$, which only contain informative samples. The reported average accuracy is averaged over 20 randomly initialized runs. For parameter setting, in RCPMM-CC, the threshold to prune noise samples is set as 0.01, and the threshold to prune cluster is set as 0.008.



(a) (b)

Fig. 4. (a) Average accuracy comparison on synthetic dataset 1 with 681 informative samples; (b) Average accuracy comparison on synthetic dataset 2 with 2579 informative samples.

Table II. The Number of Samples in Three Media Platforms in Each Time Interval

	02/04 - 03/10	03/11 - 03/17	03/18 - 03/24	03/25 - 03/31
Twitter	6, 761, 520	1, 181, 245	1, 052, 639	1, 104, 873
News	Nil	632	548	385
Flickr	Nil	2000	2000	2000

The balancing parameters are all set as 1. The numbers of clusters are set as random integers (range 0–10) larger than the ground-truths. In GS-HOCC, the numbers of clusters are set as the sum of corresponding ground-truths and outlier noise size.

Figure 4 shows the average accuracy between RCPMM-CC and GS-HOCC on both synthetic datasets. From the reported results, we have the following observations. (1) When data are clean, both RCPMM-CC and GS-HOCC achieve good performance on coclustering, that is, above 0.94 in average accuracy in two synthetic datasets. When data are mixed with outliers, the proposed RCPMM-CC performs better than GS-HOCC with all different percentage of mixtures, as GS-HOCC does not process those outliers. Moreover, when the percentage of outliers increases, the gap of average accuracies between the two methods increases. (2) When the size of dataset increases (4 times from dataset 1 to dataset 2), the average accuracy for each percentage of mixture is stable for RCPMM-CC, while that for GS-HOCC slightly decreases.

7. EXPERIMENTAL EVALUATION ON REAL WORLD DATA

In this section, we present the experimental results with real world data on emerging keyword extraction, emerging topic detection and topic elaboration respectively.

7.1. Real World Dataset

To evaluate our work, we investigate the time period from March 11th 2011 to March 31st 2011. The time interval for the detection can be set as minutes, hours, days, weeks and months. In our experiment, we set it as one week, because the ground-truths, which will be introduced in Section 7.3.1, is based on weeks. Please note that, our method can be applied on other intervals, such as month, day, and even hour, as long as sufficient data have been collected. Table II shows the number of Twitter, news samples and Flickr images in each time interval.

—*Twitter*. We collect a set of around 45,000 users, who post more than 20 tweets each week. We download all the tweets from these users with the posted time from Feb. 4th 2011 to Mar. 31st 2011. Note that the time period for data collection is five weeks longer than the targeting period to provide the historical information for aging theory. By removing the non-English tweets, we obtain around $10M$ tweets posted in eight weeks. The content, posted timestamp and user id are saved.

—*New York Times*. New York Times provides us an effective article search API to find the news published in a certain period.[7] We only download those that have high relevance with top 400 emerging keywords in Twitter. The titles, main bodies and attached images are saved.

—*Flickr*. We similarly focus on those that are relevant to the top 400 emerging Twitter keywords. The first 5 searched images, whose tags, titles or descriptions include one of top 400 emerging keywords and the shot time or uploaded time is within the corresponding time interval, are crawled.

7.2. Emerging Keyword Extraction

After necessary text preprocessing steps on Twitter data, including word separation and stopwords filtering, we compute the emerging weight $a_w^{keyword}(k)$ for w-th keyword in the k-th time interval. Table III illustrates the top 20 detected emerging keywords in each time interval. The values in parentheses are the corresponding emerging weights. The top 3 keywords for the first week are all with extremely high relevance to Japan earthquake. The fourth keyword in the first week is "Oscars", which is related to the event that "Oscar was hold in late February and early March." The fifth keyword "Rebecca" corresponds to a famous singer Rebecca Black, whose YouTube music video received over 16 million views for a short while in this week. In the second week, keyword "Rebecca" raises to the first place. The second keyword "iPad" corresponds to the release and launch of iPad2. We can see that most of the keywords are related to the hot events or topics in these weeks.

7.3. Emerging Topic Detection

Based on the extracted emerging keywords on Twitter data, RCPMM-CC is utilized to detect emerging topics on multiple platforms.

7.3.1. Baselines, Ground-truths and Criteria. We employ six baselines.

—K-means: K-means clustering [Dhillon and Modha 2001] on Twitter
—BCC-TN: Bregman Coclustering [Banerjee et al. 2007] on Twitter and News
—BCC-TF: Bregman Coclustering [Banerjee et al. 2007] on Twitter and Flickr
—ITCC-ALL: Information Theoretic Coclustering [Gao et al. 2006] on three platforms with star structure
—ROCC-TN: Robust Overlapping Coclustering [Deodhar et al. 2008] on Twitter and News
—ROCC-TF: Robust Overlapping Coclustering [Deodhar et al. 2008] on Twitter and Flickr
—RCP-CC: Robust Cross-Platform Coclustering on single modality
—GS-HOCC: Gerneral-Structured High-Order Coclustering [Bao et al. 2012]

K-means is one side clustering on single platform. This baseline is used to evaluate the performance on emerging topic detection by not introducing the complementary media platforms. BCC-TN and BCC-TF are coclustering based on only two platforms, while ROCC-TN and ROCC-TF are two platforms coclustering with noise toleration. These baselines are used to evaluate the performance on emerging topic detection with only one introduced complementary media platform. ITCC-ALL is high-order coclustering with star structure, which neglects the correlations between news and Flickr image

[7]http://developer.nytimes.com/docs/article_search_api.

Table III. The Top 20 Emerging Keywords in Each Time Interval

	03/11 - 03/17	03/18 - 03/24	03/25 - 03/31		03/11 - 03/17	03/18 - 03/24	03/25 - 03/31
1	Japan (0.2524)	Rebecca (0.0397)	Japan (0.0487)	11	Japanese (0.0268)	Austin (0.0213)	iPad (0.0180)
2	tsunami (0.1027)	iPad (0.0296)	Gaga (0.0337)	12	Patrick (0.0255)	moon (0.0176)	cricket (0.0173)
3	earthquake (0.0689)	H1N1 (0.0268)	Lady (0.0298)	13	facebook (0.0220)	April (0.0166)	earthquake (0.0173)
4	Oscars (0.0598)	Sheen (0.0264)	Jackie (0.0265)	14	Fukushima (0.0217)	March (0.0153)	good (0.0158)
5	Rebecca (0.0567)	Taylor (0.0250)	F1 (0.0241)	15	Potter (0.0214)	Oscar (0.0139)	Twitter (0.0152)
6	nuclear (0.0508)	Libya (0.0249)	Austin (0.0240)	16	Harry (0.0207)	winning (0.0138)	love (0.0140)
7	Austin (0.0503)	Charlie (0.0248)	tsunami (0.0237)	17	relief (0.0201)	Japan (0.0116)	day (0.0134)
8	Dogg (0.0407)	tsunami (0.0223)	YouTube (0.0235)	18	plant (0.0194)	apple (0.0116)	live (0.0119)
9	nate (0.0370)	mobile (0.0219)	March (0.0235)	19	st (0.0190)	love (0.0115)	apple (0.0114)
10	quake (0.0317)	Elizabeth (0.0215)	Britney (0.0187)	20	winning (0.0177)	supermoon (0.0113)	rumors (0.0110)

Table IV. Ground Truths of Top 10 Emerging Topics during 03/18/11-03/24/11

Rank	Topic	Description
1	Rebecca Black	A music video by Rebecca Black on YouTube that many people find amusing or annoying
2	Full Moon	Last Saturday's full moon appeared 14% bigger and up to 30% brighter than normal.
3	HIN1	There are 100 cases of influenza AH1N1 in US and other Latin American countries.
4	UK Comic Relief	Red Nose Day is part of the UK Comic Relief campaign, which is held each March by the BBC.
5	Knut	A Berlin zoo official says world-famous polar bear Knut has died.
6	Elizabeth	Elizabeth Taylor has died at the age 79 of heart failure.
7	Libyan Conflict	French Rafale fighter jets are flying combat missions over Libya. Missiles were fired from U.S. warships on targets in Libya. Operation Odyssey Dawn is the name for this operation.
8	Soccer /Football	Gilson Kleina will take over as interim coach. American international Stuart Holden was the victim of an awful two-footed tackle by Manchester United center back Johnny Evans that saw the defender lunge for a 50-50 ball and dive straight onto the midfielder's knee.
9	Alexz Johnson	Alexz Johnson is a Canadian singer-songwriter, and a actor in the CTV series Instant Star.
10	UFC	Ultimate Fighting Championship 128 was held with many notable matches.

samples, and does not process the mixed outlier samples. The only difference between RCP-CC with our method is that RCP-CC uses text as the single modality. This baseline is to show the necessity to handle multimodalities for better performance. GS-HOCC is our previous work, which is based on information theoretic coclustering and does not pruning away the outlier samples during coclustering. This baseline is to validate the effectiveness after de-noising.

Published by Mashable, a company focusing on engaging and analyzing the online information, "Top Twitter trends each week" [WWW b] provides the ground truths for the emerging topics on Twitter in each week. Table IV shows the ground truths of emerging topics during 03/18/11–03/24/11.

We ask 25 participants to assign the obtained top 10 emerging keyword clusters with the ranks of corresponding ground-truths. If there is no corresponding ground truths topic, we assign it with 0. For the final assignments, the most voted ones by the participants were recorded. We also ask these participants to label the top 6 emerging keywords in each topic cluster whether they are relevant to the topic or not.[8] All the participants are allowed to search the related news of that period from Google news to get the general information of all the cluster topics. If the keyword is relevant to the topic, it is assigned with 1, while if the participant thinks the keyword is not relevant to the topic, or he is not sure, it is assigned with 0.

We use three criteria on emerging topic detection evaluation: precision, normalized discount cumulative gain (NDCG) [Jarvelin and Kekalainen 2002] on topics, and mAP at 6 on keywords. Precision is utilized to evaluate the accuracy of estimated top 10 topics, NDCG is to evaluate the rank of those topics, and mAP is to evaluate the average precision of the top 6 keywords in each topic cluster.

7.3.2. Features and Parameter Settings. Each image is extracted to an 809-dimensional feature vector as the content representation [Zhu et al. 2008], including 81-dimensional

[8]In this experiment, participants do not need to check whether the topic of keywords cluster has the corresponding ground-truths topic, they only need to focus on confirming the relevance between keywords and their cluster topic.

Table V. Top 10 Emerging Topics via RCPMM-CC during 03/18/11-03/24/11

Rank	RCPMM-CC	Corresponding Ground-truth	Ground-truth Rank
1	Rebecca, black, Friday, YouTube, pop, singer, teenage, worst, annoying	Rebecca Black	1
2	nuclear, tsunami, Japan, earthquake, Fukushima, plant, safe	Nil	0
3	Sheen, Charlie, tigerblood, half, followers, actor, men	Nil	0
4	iPad, mobile, apple, app, iPhone, ios, os, mac, touch, download, line, store	Nil	0
5	H1N1, flu, spread, fever, sick, alert, chicken, hospital, die,	H1N1	3
6	Tylor, Elizabeth, actress, died, 79, heart, jewelry, beautiful, star	Elizabeth Taylor	6
7	Libya, spring, Libyan, mission, peace, conflict, attack	Libyan Conflict	7
8	moon, supermoon, full, moonlight, bright, photographer, biggest	Full Moon	2
9	Oscar, party, carpet, red, Franco, dress, golden, Halle, inception, star	Nil	0
10	panel, session, coverage, product, circles, system, interactive, pax	Nil	0

color moment, 37-dimensional edge histogram, 120-dimensional wavelet texture feature, 59-dimensional LBP feature, 512-dimensional GIST feature. The text modality is represented by TF-IDF feature.

There are two components that need to be selected in our proposed RCPMM-CC: 1) the Bregman divergence suitable for the given data matrix, 2) a coclustering basis to get the approximation of the given data matrix. In our application, the clustering between Twitter and News as well as Twitter and Flickr is both text based, while the clustering between News and Flickr involves texts and images. Thus, we choose I-divergence and $\mathcal{C}_5$ basis [Banerjee et al. 2007] as this combination is reported the best in the document clustering.

The parameters are set as follows. τ^i, τ^t, which are for balancing image and text modalities, are tuned from $(0.1, 0.9)$ by increasing 0.1 to τ^i and decreasing 0.1 to τ^t in each step. α, β, ξ are tuned from $(0.1, \frac{1}{2}(1-\alpha), \frac{1}{2}(1-\alpha))$ to $(0.9, \frac{1}{2}(1-\alpha), \frac{1}{2}(1-\alpha))$ by increasing 0.1 to α in each step. In our experiment, τ^i, τ^t are set as 0.3, 0.6. α, β, ξ are set as 0.6, 0.2, 0.2. The numbers of remaining samples in Twitter keywords, News and Flickr images are $k = 320, l = 80, f = 50$ as there are plenty of irrelevant news and images in News and Flickr datasets. The numbers of clusters of Twitter, News and Flickr are tuned from $(20, 20, 20)$ to $(70, 70, 70)$ by increasing 5 in each step. Since the computational cost of our algorithm is very low, it does not take too much time on tuning them.

7.3.3. Experimental Results. Table V illustrates the top 10 topics during 03/18/11-03/24/11 obtained by RCPMM-CC, together with the corresponding ground-truths given by Mashable, and their ranks. The first column is the rank of the obtained topics, and the second one lists the keywords in those topics. The third and last columns show the corresponding ground-truths and their ranks, which are manually labeled by participants. From this table, RCPMM-CC correctly hits five topics in the top 10 ground-truths. Besides these topics hitting the ground truths, we also examined the other topics. For example, the second cluster responds to that Japan earthquake happened in the previous week, and people were still paying attention to it for the secondary disasters and donation. The third topic responds to “Charlie Sheen,” a famous actor, and he was involved in a litigation and lost a bid to block arbitration in this week. And the fourth topic cluster is about the sale of iPad2. It is the first week for iPad sale in US, and long queues of buyers occurred at many Apple stores. Although these topics

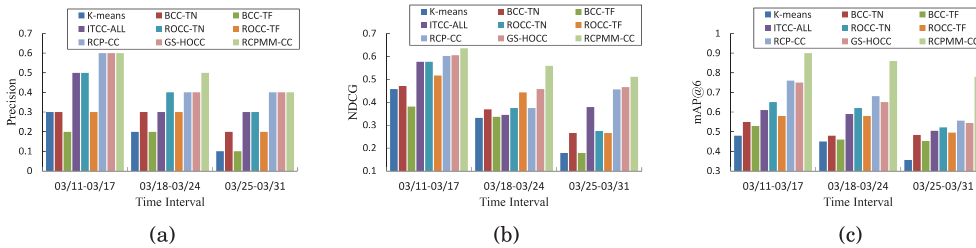


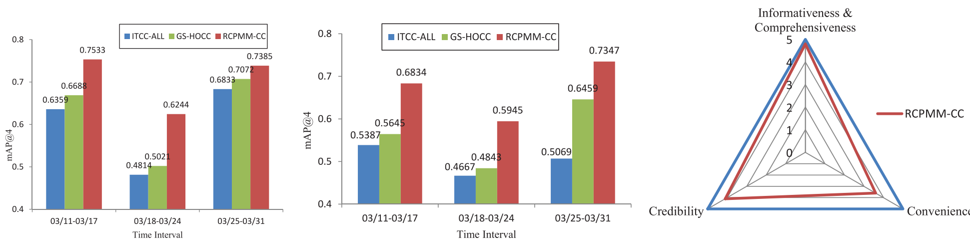
Fig. 5. (a) Precision comparison on emerging topic detection; (b) NDCG comparison on emerging topic detection; (c) mAP comparison on top 6 emerging keywords.

are not in the top 10 ground truths, comparing to those nonmatched topics in ground truths, shown in Table IV, they are still considered as trending topics at the time.

Figure 5(a) and (b) show the precisions and NDCGs on detected topics, and Figure 5(c) shows mAPs of top 6 emerging keywords in top 10 topic clusters. From the results, we have the following observations. 1) Clustering on single platform, for instance, K-means, performs comparable or worse than other methods. This shows the necessity of introducing complementary platforms. 2) Coclustering on two platforms in general performs worse than that on three platforms, eg. BCC-TN/BCC-TF vs. ITCC-ALL, ROCC-TN/ROCC-TF vs. RCP-CC/GS-HOCC/RCPMM-CC. This proves that both complementary platforms are beneficial to topic detection. 3) Coclustering on Twitter and New York Times achieves better performance than that on Twitter and Flickr, for instance, BCC-TN vs. BCC-TF, ROCC-TN vs. ROCC-TF. The reason might be that samples from New York Times include more emerging keywords than those from Flickr, thus the relevant samples from New York Times could have higher probability to cluster together, and so the better results. 4) Coclustering on single texture modality, for instance, RCP-CC, performs slightly worse than coclustering on multimodalities in Figure 5(a), but much worse in Figure 5(b) (c). The reason of the large gaps in NDCG and mAP could be that the visual modality is beneficial to discover more samples belonging to the same cluster, and this will increase the emerging weight of the cluster, as well as reveal more true corresponding keywords. 5) For coclustering on two platforms, robust methods, for instance, ROCC-TN and ROCC-FT, always achieve better than regular methods, for instance, BCC-TN and BCC-TF. This proves the advantage of processing outlier noises. However, compared with coclustering methods on three platforms, this achievement is relatively small, especially in Figure 5(c), likely due to the low precision of topic detection. 6) For coclustering in pairwise correlated structure on three platforms, RCPMM-CC is either comparable or better than GS-HOCC, which does not prune away outliers. Good results of both GS-HOCC and RCPMM-CC in Figure 5(a) demonstrate the effectiveness of coclustering in pairwise correlated structure on three platforms. While RCPMM-CC performs much better than GS-HOCC in Figure 5(b) (c), as the samples in each detected emerging topics are mostly informative. This shows that the process of discarding outliers will bring better performance on topic detection, especially on topic ranking and finding true emerging keywords for each topic.

7.4. Cross-Platform Elaboration

In this experiment, the numbers of selected news and Flickr image samples for elaboration $\hbar_n$ and $\hbar_f$ are set as 4. We choose ITCC-ALL and GS-HOCC as our baseline on all the three platforms. We asked 25 participants to manually label whether the obtained news and images are related to the corresponding Twitter keyword clusters or not. If not relevant, labeled the sample 0, otherwise, 1. We recorded the most voted ones as ground-truths. To show the effectiveness of those manually labeled data, we



(a) (b) (c)

Fig. 6. (a) mAP comparison on elaboration with news from New York Time; (b) mAP comparison on elaboration with Flickr image; (c) The user experience from three aspects.

count the samples that have been labeled the same from more than 20 participants. It shows that more than 85% samples received the overwhelming majority votes. With the ground-truths, we can calculate the mAPs on top 10 emerging topics. Figure 6(a) (b) show the results on New York Times and Flickr respectively in each time interval. We can see that RCPMM-CC always performs better than ITCC-ALL and GS-HOCC in all the three weeks on both platforms. For the same week, the increment of mAP on New York Times are mostly smaller than those on Flickr. This might be because the correlations between news and Flickr image samples, especially the relationship between the visual modality on two media platforms, could be much beneficial to cluster the related Flickr images together.

Specifically, we have the following observations on the detected emerging topic in the first week. The first emerging topic is the Japan Earthquake, which occurred in 11th March. The first ranked news reported the general information on earthquake and tsunami. The attached news photos show an area in Sendai City, which had been swept by the tsunami. The second one reported the living status of the survivors and provided another viewpoint of the Japan Earthquake. The other two selected news stones, which are not shown in this figure, are related to the nuclear crisis and donation. The Flickr images show another views of the tsunami, homes engulfed by fire and water, and Fukushima nuclear power plant. Taking the third emerging topic, “Nate Dogg died,” as another example, the selected Flickr images include not only the photo taken near his house, for instance, the first image, but also the memorial activities taking by his fans, for instance, the second image.

We also asked 25 participants to evaluate the elaboration results from three aspects: 1) informativeness and comprehensiveness; 2) convenience; 3) credibility. All the evaluations were categorized into five levels, for instance, {1, 2, 3, 4, 5}, indicating “very bad,” “bad,” “average,” “good,” and “very good.” The average statistic from all the participants are shown in Figure 6(c). The results show that users mostly think that the selected news and images are informative and comprehensive. The aspect of credibility gets “good,” while the score of convenience is between “average” and “good”. The reason for low score of convenience might be that the results are based on weeks, and users would like to receive the latest information as soon as possible. As we discussed in Section 7.1, provided sufficient data, our work can work on short time interval, like an hour, or a day.

8. CONCLUSIONS

In this article, we have proposed a framework of detection and elaboration of emerging topics on microblog type platform with samples by introducing news portals and image sharing websites as complementary platforms. Robust Cross-Platform Multimedia Coclustering is proposed to simultaneously discover the clusters in three pairwise dependent media platforms, and meanwhile prune away noninformative samples. The future work includes two aspects. Firstly, although our approach has the ability to remove the outlier data, high quality news and image collection is still desired in topic extraction and cross-platform elaboration. Thus we plan to work on keyword group generation to obtain more relevant news and images. Secondly, the proposed algorithm utilizes the between-similarities over different data platforms and we feel that it can be further extended to work on the within-similarities in each platform also.

ELECTRONIC APPENDIX

The electronic appendix to this article can be accessed in the ACM Digital Library.

REFERENCES

L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes. 2013. Sensing trending topics in Twitter. *IEEE Trans. Multimedia* 15, 6.

Foteini Alvanaki, Michel Sebastian, Krithi Ramamritham, and Gerhard Weikum. 2011. EnBlogue: Emergent topic detection in Web 2.0 streams. In *Proceedings of the ACM International Conference on Management of Data*. 1271–1274.

A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. 2007. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *J. Machine Learning Research* 8, 1919–1986.

Romil Bansal, Radhika Kumaran, Diwakar Mahajan, Arpit Khurdiya, Lipika Dey, and Hiranmay Ghosh. 2012. TWIPIX: A web magazine curated from social media. In *Proceedings of the ACM International Conference on Multimedia*. 1355–1356.

Bing-Kun Bao, Weiqing Min, Teng Li, and Changsheng Xu. 2015. Joint local and global consistency on interdocument and interword relationships for co-clustering. *IEEE Trans. Cybernetics* 45, 1, 15–28.

Bing-Kun Bao, Weiqing Min, Ke Lu, and Changsheng Xu. 2013. Social event detection with robust high-order co-clustering. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*. 135–142.

Bing-Kun Bao, Weiqing Min, Jitao Sang, and Changsheng Xu. 2012. Multimedia news digger on emerging topics from social streams. In *Proceedings of the 20th ACM International Conference on Multimedia*. 1357–1358.

R. Cai, L. Lu, and A. Hanjalic. 2008. Co-clustering for auditory scene categorization. *IEEE Trans. Multimedia* 10, 4, 596–606.

C. Chen, Y. T. Chen, Y. Sun, and M. Chen. 2003. Life cycle modeling of news events using aging theory. In *Proceedings of the International Conference on Machine Learning*. 47–59.

K. Y. Chen, L. Luesukprasert, and S. Chou. 2007. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Trans. Knowl. Data Eng.* 19, 8, 1016–1025.

M. Deodhar, H. Cho, G. Gupta, J. Ghosh, and I. Dhillon. 2008. Robust overlapping co-clustering. Tech Rep. IDEAL-TR09, Department of ECE, University of Texas at Austin.

I.S. Dhillon and D.S. Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine Learning* 42, 1, 143–175.

I. S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. 269–274.

I. S. Dhillon, S. Mallela, and D. S. Modha. 2003. Information-theoretic co-clustering. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*.

Chris Ding, Tao Li, Wei Peng, and Haesun Park. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*. 126–135.

M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. König. 2008. Blews: Using blogs to provide context for news articles. In *Proceedings of the AAAI Conference on Weblogs and Social Media*. American Association for Artificial Intelligence.

B. Gao, T. Y. Liu, and W. Y. Ma. 2006. Star-structured high-order heterogeneous data co-clustering based on consistent information theory. In *Proceedings of the International Conference on Data Mining*. IEEE, 880–884.

G. Greco, A. Guzzo, and L. Pontieri. 2010. Coclustering multiple heterogeneous domains: Linear combinations and agreements. *IEEE Trans. Knowl. Data Eng.* 22, 12, 1649–1663.

K. Jarvelin and J. Kekalainen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4, 422–446.

Stefanie Jegelka, Suvrit Sra, and Arindam Banerjee. 2009. Approximation algorithms for tensor clustering. *Algorithmic Learning Theory*, Lecture Notes in Computer Science, vol. 5809, 368–383.

Shiva Prasad Kasiviswanathan, Prem Melville, Arindam Banerjee, and Vikas Sindhwani. 2011. Emerging topic detection using dictionary learning. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 745–754.

R. E. Kass and L. Wasserman. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.*, 928–934.

T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan. 2015. Crowded scene analysis: A survey. *IEEE Trans. Circuits Syst. Video Technol.*

Bo Long, Zhongfei Mark Zhang, and Philip S. Yu. 2005. Co-clustering by block value decomposition. In *Proceedings of the ACM International Conference on Knowledge Discovery in Data Mining*. 635–640.

Miles Osborne, Sasa Petrovic, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2012. Bieber no more: First story detection using Twitter andWikipedia. In *Proceedings of the SIGIR Workshop on Time-Aware Information Access*.

D. Pelleg and A. Moore. 2000. X-means: Extending K-means with efficient estimation of the number of clusters. In *Proceedings of the 17th International Conference on Machine Learning*. 727–734.

S. D. Roy, T. Mei, W. Zeng, and S. Li. 2012. Empowering cross-domain internet media with real-time topic learning from social streams. In *Proceedings of the IEEE International Conference on Multimedia and Expo*.

T. Sakaki, M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*. 851–860.

Jitao Sang, Changsheng Xu, and Jing Liu. 2012. User-aware image tag refinement via ternary semantic analysis. *IEEE Trans. Multimedia* 14, 3, 883–895.

Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. 2009. Event detection and tracking in social streams. In *Proceedings of the International Conference on Weblogs and Social Media*.

Giuseppe Serra, Thomas Alisi, Marco Bertini, Lamberto Ballan, Alberto Del Bimbo, Laurent Walter Goix, and Carlo Alberto Licciardi. 2013. Demo paper: Stamat: A framework for social topics and media analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops*. 1–2.

Y. Takama, A. Matsumura, and T. Kajinami. 2006. Visualization of news distribution in blog space. In *Proceedings of the IEEE International Conference on Web Intelligence and Intelligent Agent Technology*. 413–416.

S. Tan, C. W. Ngo, H. K. Tan, and L. Pang. 2011. Cross media hyperlinking for search topic browsing. In *Proceedings of the ACM International Conference on Multimedia*. 243–252.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 267–273.

Jianke Zhu, Steven C. H. Hoi, Michael R. Lyu, and Shuicheng Yan. 2008. Near-duplicate keyframe retrieval by nonrigid image matching. In *Proceedings of the 16th ACM International Conference on Multimedia*. 41–50.

Received January 2014; revised August 2014, January 2015; accepted January 2015