Robust Latent Poisson Deconvolution From Multiple Features for Web Topic Detection

Junbiao Pang, Fei Tao, Chunjie Zhang, Weigang Zhang, Qingming Huang, *Senior Member, IEEE*, and Baocai Yin, *Member, IEEE*

***Abstract*—Detecting "hot" topics from the enormous user-generated content (UGC) data on web poses two main difficulties that the conventional approaches can barely handle: 1) poor feature representations from noisy images or short texts, and 2) uncertain roles of modalities where the visual content is either highly or weakly relevant to the textual cues due to the less-constrained UGC. In this paper, following the detection-by-ranking approach, we address above challenges by learning a robust latent representation from multiple, noisy and a high probability of the complementary features. Both the textual features and the visual ones are encoded into a k-nearest neighbor hybrid similarity graph (HSG), where nonnegative matrix factorization using random walk is introduced to generate topic candidates. An efficient fusion of multiple HSGs is then done by a latent poisson deconvolution, which consists of a poisson deconvolution with sparse basis similarity for each edge. Experiments show significantly improved accuracy of the proposed approach in comparison with the state-of-the-art methods on two public datasets.**

***Index Terms*—*K*-nearest neighbor similarity graph, latent poisson deconvolution (LPD), multi-view learning (MVL), user-generated content (UGC), web topic detection.**

I. Introduction

WITH the rapid development of social media, User-Generated Content (UGC) [28] is quite pervasive for people to either share or exchange their options and experiences. Meanwhile, the content of UGC is more sparse, unconstrained and less predicable than that of the professionally edited articles, since everybody is both the producer and the consumer of media. As a result, the unprecedented explosion in the volume of "we-media" has made it difficult for web users to quickly access hot and interesting topics [32]. Web topic detection [28], [42] is such an effort to organize web data into more meaningful and interpretable topics automatically. Web detecting topics here is defined as the task of discovering of a tiny fraction of webpages strongly connected by a seminal event from a large amount of social media [28].

One of the important approaches is to exploit multiple modalities of data themselves. Generally speaking, "we-media" is often posted at will across multiple modalities, reflecting social realities from multiple aspects. Therefore, these less-constrained UGC data often face several challenging problems: 1) the probability of a deficiency of some modality; 2) inefficient feature representations either from short text [43] or noisy images; and 3) the uncertain roles of different modalities. The last is an extremely universal phenomenon in social media, where the visual cues are possibly more important than the textual ones to express the content of a webpage while the textual cues may serve a dominant role for other webpages. The existing methods, at the modality level, combine multiple representations from the different modalities with possible noises. For instance, different modalities are linearly averaged into a unified representation by the well-tuned weights [42]. However, the varying roles of different modalities in each sample are not considered.

Therefore, we seek a *robust*, and *datum-wise* framework to exploit *multiple* and *noisy* feature representations, based on two motivations. First, although an enormous volume of literatures has been devoted to the feature fusion problem, most of them consensually assume that features nearly have no noise. Second, we want to avoid the disadvantages of the popular modality-level fusion methods, e.g., the linear weight approach [42]. The modality-level methods obviously have a difficulty in dealing with the uncertain role of different modalities. In summary, our goal is to robustly detect topics from multiple noise representations of the multi-media data where the importance of the modality varies with respect to each individual webpage.

In this paper, we propose a latent Poisson deconvolution (LPD) framework to explicitly handle the possible noises associated with the different feature representations. As shown in Fig. 1, following similarity cascade (SC) [28], we in the preprocessing stage extract multi-view features from the different modalities, and then compute a similarity graph with each descriptor. The adverse impacts of noises are naturally encoded

Manuscript received April 11, 2016; revised August 1, 2016; accepted August 1, 2016. Date of publication August 5, 2016; date of current version November 15, 2016. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2012CB316400 and Grant 2015CB3351800, in part by the Natural Science Foundation of China under Grant 61332016, Grant 61472387, Grant 61303153, Grant 61390510, and Grant 61303154, in part by the Beijing Post-Doctoral Research Foundation, in part the Beijing Municipal Commission of Education under Grant KM201610005034, and in part by the Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jing-Ming Guo. *(Corresponding author: Weigang Zhang.)*

J. Pang is with the Beijing Key Laboratory of Multimedia and Intelligent Software Technology, College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China (e-mail: junbiao_pang@bjut.edu.cn).

F. Tao and C. Zhang are with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: ftao@jdl.ac.cn; cjzhang@jdl.ac.cn).

W. Zhang is with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, China, and also with the University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100049, China (e-mail: wgzhang@hit.edu.cn).

Q. Huang is with the University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100049, China, and also with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: qmhuang@jdl.ac.cn).

B. Yin is with the Advanced Invocation Center for Future Internet Technology, Dalian University of Technology, Dalian 116024, China, and also with Beijing University of Technology, Beijing 100124, China (e-mail: ybc@dlut.edu.cn).

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.2016.2598439

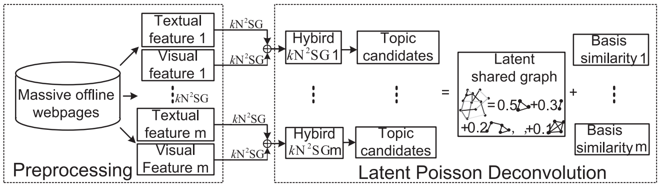


Fig. 1. Proposed LPD framework.

into the similarity graph. To partially reduce these unfavorable impacts, we only select the top-k most similar values to build a k-Nearest Neighbor Similarity Graph (k-N^{2}SG). Next, each paired k-N^{2}SGs from both the textual modality and the visual one are equally weighted into a Hybrid Similarity Graph (HSG). Instead of directly using k-N^{2}SG from each modality, the motivation behind HSGs is that the visual cues still contain some minor yet meaningful information, although it is extremely difficult to convert the visual cues into the social-related concepts due to the "semantic gap" challenge.

Further in generating topic candidates from these HSGs, nonnegative matrix factorization using random walk (NMFR) [40] is used to generate multi-granularity candidates. The advantage lies in that random walk empowers NMF to capture manifold structure among data points, producing high-quality topic candidates.

In the detection-by-ranking stage [28], rather than firstly learning a shared graph (which is a common approach in multiview learning, however unnecessary in our approach as will be discussed) we instead propose LPD to learn the reconstructed latent graph and sparse basis similarities from different HSGs, in the hope of not only avoiding the complicated optimization problem but also adaptively fusing multiple cues in the datumwise fashion. The last is crucial to deal with the uncertain roles of the different modalities for individual webpage.

By considering the uncertain roles of the different modalities, LDP learns a latent shared representation for each webpage. To the best of our knowledge, this is the first to investigate multiple-view learning (MVL) for web topic detection, by adaptively fusing features from the multiple modalities int the datum-wise fashion. The proposed method is conceptually simple, yet exceptionally powerful. We develop a web topic detection method that exceeds five state-of-the-art approaches [10], [13], [18], [28], [42] on two public datasets.

The rest of this paper is organized as follows. Section II reviews the related work. We describe the details of our approach in Section III. Experimental results are presented in Section V and the paper is concluded in Section VI.

II. Related Work

Topic detection from the single modality of data: In the single-modality based approach, it is reasonable to assume that one of the modalities always plays the dominant role in determining the content of social media. Based on this assumption, existing approaches define different topic patterns either from the textual modality or from the visual one. Due to the "semantic gap" between visual information and the social concepts, nearly all approaches only use the textual modality, assuming that elements in a topic have higher similarities between each other, e.g., news [2], [3], [9], blogs [33]. Therefore, many literatures about topic detection consider web topics as clusterings. There are three important research threads in the topic-as-clustering method.

One popular way explicitly defines an intra-similarity to group a set of elements into a topic. For example, Yang *et al.* [39] propose a classical framework in which the group-average clustering is used to discover topics. In [35], an agglomerative clustering method based on average pair-wise similarities is proposed to group news into topics. He *et al.* [17] propose periodic, aperiodic features and the characteristics of word trajectory, for event detection in news.

The second method adopts topic models to handle the polysemous phenomenon in topic detection [28]. Topic models have been proposed to infer hidden themes for document analysis, allocating a document into several hidden topics. The classical topic models include latent dirichlet allocation (LDA) [6], hierarchical dirichlet processes (HDP) [34], probabilistic latent semantic analysis (pLSA) [19] and various variations. In [38], nonnegative matrix factorization (NMF) shows more accurate performance than that of the spectral methods in document clustering. These topic models generally work well on long and structured documents [16]. Compared with the intra-similarity approach, the second method tends to fail on short and noisy text from social media since they are heavily dependent on word co-occurrence. Besides, the assumption adopted in topic models, every webpage evolves into a topic, does not hold for social media. Because web topic is a kind of a needle in a haystack, discovering a subset of meaningful and interesting webpages from an enormous web data.

The third approach incorporates the possible side information that could probably guide the text-based clusterings. This category aims to utilize possible cues from the *other* information channels beyond itself. For instance, [33] proposes to use queries recorded in searching engines to filter out false positives. Similarly, [13] detects topics in a user-oriented manner and proposes a query-guided topic detection method. In [25], by leveraging the external sources such as online news and blogs, news videos are clustered into a hierarchical structure. Often the most severe drawback of this approach is that the quality of side information should be close to that of the supervised one.

Generally speaking, the topic-as-clustering method works well in some scenarios, e.g., news, blogs and scientific documents [23]. As mentioned above, however, web topics are not equivalent to these long and structured documents since the noisy and spare data are more serious than that of other applications. Noticing above challenges in the topic-as-clustering method, the detecting-by-ranking approach is proposed to covert the topic detection as ranking problem [28]. More concretely, the multi-granularity topic candidates are firstly generated to describe overlapping, mutual exclusion, and subsumption relationship between topics, and then the Poisson Deconvolution (PD) [28] is used to rank the interestingness of candidates. The proposed method in this paper belongs to the PD based method. However, since our goal is to exploit the multi-modalities of data, our formulation introduces a latent shared representation, resulting in a efficient datum-wise feature fusion scheme.

Topic detection from the multi-modality of data: This approach aims to exploit the possible complementary modalities from data themselves. In this approach, there are two important threads of research. One extends clustering algorithms into the multi-modality data [7], [30], and the other is the fused similarity graph method [29], a work based on the multi-modalities fusion.

In the former case, the discovery of topics involves extending the single-modality based models into multi-modal data. For instance, multi-modal LDA [30] is proposed to group images with tags into topics. In the similarity graph method, multi-modal information is fused into the edges of a similarity graph. Cao *et al.* [10] first generate events on video tags by k-means, and then link these ones into topics based on the textual-visual similarity. For instance, Wu *et al.* [36] use weighted similarity between nearly-duplicated keyframe (NDK) and the speech transcripts from news videos.

The key problem of the fused similarity graph method is how to efficiently fuse multiple cues. Generally, the fusion problem is conceptually formulated as follows: assume that we have a set of graphs G_i extracted from different features, a fused graph G is linearly weighed as

$$G = \sum_i \alpha_i G_i \quad \text{with} \quad \alpha_i \geq 0, \text{ and } \sum_i \alpha_i = 1. \tag{1}$$

Compared with the method extending topic modelings [6] into multi-modal data, the fused similarity graph method is both computationally simple, and easily extendable for the other graph-based algorithms [1]. Although different methods are proposed to learn these weights, the fusion scheme in (1) barely considers the uncertain roles of different modalities of data; besides, the scheme (1) nearly has no ability to deal with the noises in the less-constrained UGC. In contrast, our method does not necessarily aim at designing a perfect weight scheme to fuse heterogenous graphs at the modality level, but rather adaptively fuses multiple similarities in the datum-wise fashion.

Multi-view learning: MVL is the problem of machine learning from data represented by multiple feature sets. In this paper, multiple views mean HSGs from a given dataset. Many methods have been proposed for multi-view classification [46], retrieval [20], clustering [5].

Conceptually, existing methods for MVL can be roughly categorized into two categories. The methods in the first stream integrate multi-view features into some common representation [21], [44]. For example, Bickel *et al.* [5] incorporate multi-view features to construct the loss function for clustering. The methods in the second stream project each view of features onto a common low-dimensional subspace. A representative method in this stream is canonical correlation analysis (CCA) for MVL [11]. For more recent progress on MVL, we refer the interested readers to a literature survey of MVL [37]. The proposed method in this paper belongs to the first stream. However, the common representation in our LPD is hidden, not be explicitly computed.

Recently, there has been a trend of explicitly handling the noises in multiple inputs via sparse decomposition in machine learning. For example, the robust data fusion methods [27], [41] separate the considerable noise in multiple inputs via low-rank and sparse decomposition. All have shown increased robust power of the models. Following this trend, LDP assumes that each HSG is explicitly corrupted by noises, and then learns a latent shared graph by exploiting the joint statistics.

III. Generating Candidates on Hybrid Similarity Graph

A. Combining k-N^{2}SGs Into an HSG

Given a set of data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where each sample $\mathbf{x}_i = (x_i^v, x_i^t)$ contains the visual modality x_i^v and the textual one x_i^t. We extract the multiple visual descriptors f_v^m $(m = 1, \ldots, M)$ and the multiple textual ones f_t^m $(m = 1, \ldots, M)$ from a webpage, respectively. Based on these features, we can construct a pair-wise similarity matrix S_v^m and $S_t^m, m = 1, \ldots, M$. Let s_{ij} $(s_{ij} \in S)$ denote the similarities on a pair of data points $\mathbf{x}_i$ and $\mathbf{x}_j$.[1]

Further, let (V, E, W^m) be a weighted graph with the vertex set V, the edge set E, and the corresponding weight matrix W^m, where each vertex v_i $(v_i \in V)$ associates with a webpage $\mathbf{x}_i$, and an edge e_{ij} $(e_{ij} \in E)$ associates with a weight w_{ij}^m $(w_{ij}^m \in W^m)$ between $\mathbf{x}_i$ and $\mathbf{x}_j$. We propose to use Gaussian kernels to covert the similarity matrix S^m, *i.e.*, $w_{ij}^m = \exp\left(-\|s_{ij}^m\|^2/\sigma^2\right)$ where $\|\cdot\|^2$ denotes the ℓ_2 norm and σ^2 denotes the deviation. Gaussian kernel nonlinearly scales different similarity S_{ij}^m into a uniformly one ranging from 0 to 1. Because different features from diverse modalities not necessarily use the same similarity measurement, e.g., cosine distance for TF-IDF, and Euclidean distance for fisher vector (FV) [31].

Once similarity graphs are computed, the top-k most similar data $\mathbf{x}_i$ are inserted as its neighbors on the graph, and the other similarities are assigned with zeros. In general, the selection of k is determined by the degree of noises in different datasets. The higher noise is in a dataset, the lower value is assigned to k. Because a smaller k tends to remove more incorrect or unnecessary correlations between webpages. Therefore, the k of the visual modality is usually smaller than that of the textual one. Because mining the social-related semantics from the visual information is more challenge than that of the textual one.

We term the resulting sparse graphs as k-N^{2}SG. Subsequently, a pair of the visual k-N^{2}SG (V, E_v^m, W_v^m) and the textual one (V, E_t^m, W_t^m), are equally merged into a HSG

$$G^m = (V, E^m, W^m), \quad m = 1, \ldots, M \tag{2}$$

where $E^m = E_v^m \cup E_t^m$, and $W^m = (W_v^m + W_t^m)/2$. The unbiased weight scheme in HSG (2) is a reasonable intermediate solution during the datum-wise fusion in Section IV. Moreover, (2) naturally solves the deficiency of a modality problem [29]. We summarize some notations used in this paper in Table I.

[1] In the following, we ignore the superscript and subscript in the different context, if it does not cause confusion.

TABLE I
SOME NOTATIONS IN THIS PAPER

Notation	Definition
$\mathbf{x}_i$	A multimodal sample
$w_{ij} \in W_{ij}$	A weight between $\mathbf{x}_i$ and $\mathbf{x}_j$ in graph G
$e_{ij} \in E_{ij}$	An edge in graph G
C_k	A topic candidate
$\mu_k \in \boldsymbol{\mu}$	The weight of the candidate C_k
B^m	The basis similarity matrix
G^m	A hybrid similarity graph
$\top$	The matrix transpose operator

B. Generating Topic Candidates

NMFR is carefully chosen to generate topic candidates, in order to exploit the non-neighborhood relationship on these sparse HSGs G^m. Let $U^m \in \mathbb{R}^{N \times K}$ be nonnegative and orthogonality matrix, the objective function of NMFR is as

$$\min_{U^m \geq 0} -\mathrm{Tr}({U^m}^\top A U^m) + \lambda \|U^m\|_F^2$$
$$\text{s.t. :} \;\; {U^m}^\top U^m = 1$$

in which A is the random walk distance, $A = (I - \alpha {D^m}^{-1/2} W^m {D^m}^{-1/2})^{-1}$, where $\alpha \in (0, 1)$ is a decay parameter, and D^m is a diagonal matrix with $D^m_{ii} = \sum_{j=1}^N W^m_{ij}$. [40] proposes a relaxed algorithm to optimize U^m without explicitly computing the matrix A.

By the winner-take-all principle, U^m generates the topic candidates C^m_k $(k = 1, \ldots, K)$ [28]. Formally, $C^m_k = {c^m_k}^\top \circ c^m_k$ where the indicator vector $c^m_k \in \{0, 1\}^{1 \times N}$, in each of whose bin 1 or 0 means that the candidates C^m_k whether contains the sample $\mathbf{x}_i$ or not. The operation $\circ$ means that the diagonal of matrix ${c^m_k}^\top c^m_k$ is set to zero.

IV. ROBUST LATENT POISSON DECONVOLUTION

A. Latent Poisson Deconvolution

The basic assumptions in LPD are threefolds: 1) in the context of MVL, each HSG G^m is considered as a view of a dataset, being sufficient to discover most of correlations among webpages; 2) the matrix W^m of each graph G^m is corrupted by noises; and 3) different HSGs G^m have different basis similarities to indicate the "background" similarities. Based on above assumptions, the similarity matrices W^m $(m = 1, \ldots, M)$, can be naturally decomposed into three parts as follows:

$$\forall m, \;\; W^m = W + B^m + noise \tag{3}$$

where W is a latent shared similarity matrix that reflects the underlying true correlation among webpages, B^m represents different basis similarities, and $noise$ is a noise term.

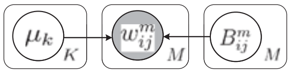


Fig. 2. Graphical model of the LPD model.

Compared with the linear weight scheme in (1), the datum-wise fusion approach in (3) instead of directly learns a fused representation W for each sample. Meanwhile, the basis similarities are optimized to discover these "background" similarities. Therefore, the datum-wise approach in (3) robustly and adaptively fuses multiple inputs. A key question arising here is how to solve the latent matrix W, and the basis similarity matrix B^m.

For the latent matrix, a naïve approach is to simply feed the matrix W into PD [28], once the latent matrix W is computed. Following Poisson noise assumption adopted in PD [28], this idea has to apply twice expectation-maximization (EM) algorithm to separatively optimize W and $\boldsymbol{\mu}$, resulting in an inefficient numerical solution. More efficiently, we can approximate $W \approx \sum_{k=1}^K \mu_k C_k$ where $C_k \in \text{union}(C^m_k)$, $(m = 1, \ldots, M$, $k = 1, \ldots, K)$. Because the latter approach only applies once EM algorithm.

For the basis similarity, each basis similarity matrix B^m represents the difference between W and W^m. Since we assume each graph G^m is sufficient to identify most of the shared graph G, it is reasonable to assume that only a small fraction of elements in W^m being significantly different from the corresponding ones in W. That is, basis similarity matrix B^m tends to be sparse.

Under the above assumptions, the similarity of an edge can be decomposed as follows:

$$\forall m, \;\; w^m_{ij} \sim \text{Poisson}\,(a_{ij}) \tag{4}$$

where $a_{ij} = \sum_{k=1}^K \mu_k C_{k_{ij}} + B^m_{ij}$. This leads to a graphical representation of the model in Fig. 2.

Given the model parameters, we formulate LPD as the MVL problem by minimizing the following regularized log-likelihood $\mathcal{L}(\boldsymbol{\mu}, B^m)$:

$$\min \underbrace{- \ln \prod_{m=1}^M \prod_{w^m_{ij} \in W^m} \frac{{a_{ij}}^{w^m_{ij}} e^{-a_{ij}}}{w^m_{ij}!}}_{\mathcal{L}(\boldsymbol{\mu}, B^m)} + \lambda \sum_{i=1}^M \|B^m\|_1$$
$$\text{s.t. :} \;\; \mu_k \geq 0, \;\; k = 1, \ldots, K \tag{5}$$

where λ is a non-negative trade-off parameter, $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]^\top$, and the ℓ_1 norm $\|\cdot\|_1$ is well-known to produce a sparse solution. The constraints $\mu_k \geq 0$ lead to a nearly sparse solution, automatically reducing the redundancy among the topic candidates. $\mathcal{L}(\boldsymbol{\mu}, B^m)$ in (5) learns a reconstructed, shared and latent graph from M hybrid ones in the context of MVL.

B. Optimization

The optimization problem in (5) is challenging due to its non-smooth terms $\|B^m\|_1$. We apply the idea of alternating direction method of multipliers (ADMM) [8] to convert the optimization problem into several sub-problems by introducing auxiliary variables Z^m, $m = 1, \ldots, M$

$$\min_{\boldsymbol{\mu} \geq 0, B^m, Z^m} \mathcal{L}(\boldsymbol{\mu}, B^m) + \lambda \sum_{m=1}^M \|B^m\|_1$$
$$\text{s.t. :} \; \boldsymbol{\mu} \geq 0, B^m = Z^m, \;\; m = 1, \ldots M. \tag{6}$$

In ADMM, we optimize the augmented Lagrangian of the above problem that can be formulated as follows:

$$\mathcal{L}_\rho = \mathcal{L}(\boldsymbol{\mu}, B^m) + \lambda \sum_{m=1}^{M} \|Z^m\|_1 + \frac{\rho}{2} \sum_{m=1}^{M} \|B^m - Z^m\|_F^2 + \sum_{m=1}^{M} \text{trace}({Y^m}^\top (B^m - Z^m)) \tag{7}$$

where $\rho \geq 0$ is called the penalty parameter, $\|\cdot\|_F$ denotes the Frobenius norm, and the matrices Y^m are the dual variables associated with the constraints, $B^m = Z^m, m = 1, \ldots, M$. The algorithm for solving the above augmented Lagrangian problem involves the following iterative steps:

$$\begin{aligned} \boldsymbol{\mu}^{t+1}, {B^m}^{t+1} &= \arg \min_{\boldsymbol{\mu} \geq 0, B^m} \mathcal{L}_\rho(\boldsymbol{\mu}, B^m, {Z^m}^t, {Y^m}^t) \\ {Z^m}^{t+1} &= \arg \min_{Z^m} L_\rho(\boldsymbol{\mu}^{t+1}, {B^m}^{t+1}, Z^m, {Y^m}^t) \\ {Y^m}^{t+1} &= {Y^m}^t + \rho({B^m}^{t+1} - {Z^m}^{t+1}). \end{aligned} \tag{8}$$

The advantage of the sequential update is that we separate multiple variables and thus optimize them at a time.

Solving for $\boldsymbol{\mu}$ and B^m: When solving for $\boldsymbol{\mu}$ and B^m in (7), the relevant terms from $\mathcal{L}_\rho$ are

$$\arg \min_{\boldsymbol{\mu} \geq 0, B^m} \mathcal{L}(\boldsymbol{\mu}, B^m) + \sum_{m=1}^{M} \text{trace}({Y^m}^\top (B^m - Z^m)) + \frac{\rho}{2} \sum_{m=1}^{M} \|B^m - Z^m\|_F^2. \tag{9}$$

By Jansen's inequality, (9) has the following upper bound:

$$\begin{aligned} &- \sum_{m=1}^{M} \sum_{w_{ij}^m \in G^m} \Bigg(w_{ij}^m \Bigg(\sum_{k=1}^{K} P_k \ln \frac{\mu_k C_{k_{ij}}}{P_k} + P_{ij}^m \ln \frac{B_{ij}^m}{P_{ij}^m} \Bigg) \\ &- \sum_{k=1}^{K} \mu_k C_{k_{ij}} - B_{ij}^m \Big) + \text{trace} \Big({Y^m}^\top (B^m - Z^m) \Big) \\ &+ \frac{\rho}{2} \|B^m - Z^m\|_F^2 \end{aligned}$$

where P_k and P_{ij}^m are the hidden variables that satisfy $P_{ij}^m + \sum_{k=1}^{K} P_k = 1$ ($P_{ij}^m \geq 0$, $P_k \geq 0$)

$$P_k = \frac{\mu_k C_{k_{ij}}}{B_{ij}^m + \sum_{k=1}^{K} \mu_k C_{k_{ij}}} \tag{10}$$

$$P_{ij}^m = \frac{B_{ij}^m}{B_{ij}^m + \sum_{k=1}^{K} \mu_k C_{k_{ij}}}. \tag{11}$$

Algorithm 1: Optimizing LPD with ADMM.

```
Input: HSGs $G^m$ $(m = 1, \ldots, M)$, Topic candidates $C_k$,
       Iteration number $T$, and Parameter $\lambda$;
Initialize: $\mu_k \leftarrow 1$ $(k = 1, \ldots, K)$, $B_{ij}^m \leftarrow 1$, and $Z_{ij}^m \leftarrow 0$
$(m = 1, \ldots, M)$, $\rho$=1.9;
for $t$ = 1 to $T$ do
    while no convergence do
        Update $\mu_k^{t+1}$, ${B_{ij}^m}^{t+1}$ by (13) and (12), respectively;
    end
    Update ${Z_{ij}^m}^{t+1}$, ${Y_{ij}^m}^{t+1}$ by (14) and (8), respectively;
end
Output: $\mu_k$ $(k = 1, \ldots, K)$, and $B^m$ $(m = 1, \ldots, M)$
```

Solving (9) by minimizing the upperbound of (10) has the closed form solution, and the non-negativity constraints are automatically taken care of

$${B_{ij}^m}^{t+1} = (-A + \sqrt{A^2 + 4\rho D})/2\rho \tag{12}$$

$$\mu_k^{t+1} = \frac{\sum_{m=1}^{M} \sum_{e_{ij} \in G^m} w_{ij}^m P_k}{M \sum_{e_{ij} \in G^m} C_{k_{ij}}} \tag{13}$$

with

$$\begin{aligned} A &= Y_{ij}^m - \rho Z_{ij}^m + 1_{e_{ij} \in G^m} \\ D &= w_{ij}^m P_{ij}^m \end{aligned}$$

where the operation $1_{e_{ij} \in G^m}$ means that if an edge e_{ij} exists in the HSG G^m, it outputs 1; otherwise, it returns 0.

Solving for Z^m: The optimization problem for Z^m can be equivalently written as

$$\min_{Z^m} \lambda \|Z^m\|_1 + \frac{\rho}{2} \|Z^m - Y^m/\rho - B^m\|_F^2$$

which has a closed form solution

$${Z^m}^{t+1} = \mathcal{S}_{\lambda/\rho}(Y^m/\rho - B^m) \tag{14}$$

where $\mathcal{S}_\alpha(\mathbf{x}) = \max(\mathbf{x} - \alpha, 0) + \min(\mathbf{x} + \alpha, 0)$ is the shrinkage operator [22].

Since the objective (5) is convex subject to nonnegative constraints, and all of its subproblems can be solved exactly based on the existing theoretical results [24], LPD converges to global optima. The optimization of LPD is summarized in Algorithm 1.

Once μ_k and B^m are computed, the interestingness of topics are ranked as, $i_k = \mu_k \cdot |C_k|$, where $|C_k|$ is the number of webpages in a topic C_k [28]. Note that our method during the evaluation adopts the Non-Maximal Suppression (NMS) [15] to handle the problem that which one is selected as the real topic if several topics intersect with each others.

C. Time Complexity

Solving $\boldsymbol{\mu}$ in (13) involves point-wise multiplication between the latent variables P_k and w_{ij}^m. For Solving for μ, and B^m, this leads to a complexity of $O(M \cdot |\mu| \cdot |C_k^m| \cdot s)$ for μ, and $O(M \cdot |W^m| \cdot s)$ for B^m, where s is the number of MM iterations, $|C_k^m|$ and $|W^m|$ are the number of edges in the topic C_k^m and in the graph G^m respectively. The time cost for Solving for Z^m can be omitted since it is usually much smaller than the update of μ and B^m. Usually, the number of the topics $|\mu|$ is larger than that of edges in a graph, $|G^m|$. Therefore, the time cost

TABLE II
SUMMARY OF DATASETS USED IN THE EXPERIMENTS

Dataset	#Topic	#Webpage	#Webpage in all topics	Dictionary size	Average #word/page	# images in dataset	Average # image/page	Comments (the cues used in our experiments are indicated in bold.)
MCG-WEBV	73	3,660	832	9,212	35	108,925	29.8	**Keyframes of video clips** and their surrounding **titles, tags and descriptions** on *Youtube* from Dec 2008 to Feb 2009.
YKS	298	8,660	990	80,294	228	71,063	8.2	**News articles on *Sina*, titles, tags, descriptions** and **keyframes of web videos** on *YouKu* from May 2012 to June 2012.

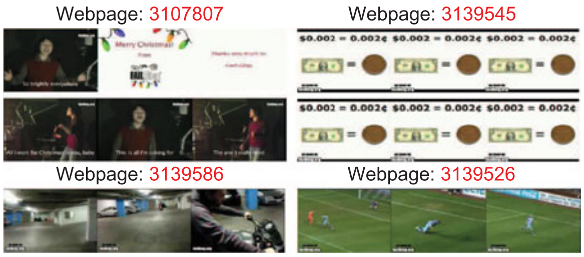


Fig. 3. Sample keyframes from some topic in MCG-WEBV. This topic contains several unexpected yet funny "fail" stories in daily life. 3107807: "Xmax fail song is an awesome win :)", 3139545: "Verizon math fail", 3139526: "Celebration Fail", and 3139586: "Motorcycle fail".

of LPD is dominated by the size of μ. The time complexity of LPD is $O(M \cdot |\mu| \cdot |C_k^m| \cdot s \cdot T)$. In practice, s is frequently smaller than 10, the maximum size of topics $|C_k^m|$ is also smaller than 100, and the number of iteration is usually less than 100. Therefore, the proposed LPD is quite efficient.

V. EXPERIMENT AND DISCUSSION

A. *Datasets, Features, Evaluations, and Experimental Setup*

We evaluate our method on two public datasets, i.e., MCG-WEBV [10] and YKS [42]. MCG-WEBV is downloaded from the "Most viewed" videos of "This month" on YouTube. YKS is a cross-media dataset crawled from YouKu and Sina respectively. The statistics of two datasets are summarized in Table II.

Since dictionaries of these sets contain multi-language words, as well as user-defined abbreviations, the dictionary size is extremely large, especially for YKS dataset. As a result, the text from social media is shorter and noisier than news articles [43]. Moreover, the visual contents of keyframes in a topic are very diverse, and contain a certain amount of noises (see examples in Fig. 3). Naturally, booting the detection performances via the noisy visual cues is a challenging task.

We choose two multi-view features for the textual cues, i.e., LDA [6] and TF-IDF, and use FV [31] for the visual cues. In our experiments, the dictionary size of LDA is 1,000. FV with 256 Gaussian components is used to represent keyframes of a video clip, where SIFT points are densely sampled from 24 × 24 image patches. Once keyframes are encoded by FV, video signature [14] is computed as similarity between two clips. The cosine distance is used to measure the similarity between textual features. For MCG-WEBV, the surrounding text of each video is considered as a set of words. While YKS in the pre-processing stage, is tokenized by *NLTK* package.[2] Therefore, two HSGs, $TF{-}IDF$+FV and LDA+FV, are generated.

As listed in Table II, the dictionary size of YKS is larger than that of MCG-WEBV. Moreover, all texts of MCG-WEBV are from titles, tags and descriptions; while, YKS, a cross-platform dataset, is a mixture of the long and the short texts. Therefore, the resulting text descriptors from YKS contain more noises than that of MCG-WEBV. Following the principle in Section III-A, a smaller k is assigned to YKS, in contrast to MCG-WEBV. In our experiments, k is assigned 100 for MCG-WEBV and 20 for YKS in the textual k-N^{2}SG. For the visual graphs, k is assigned 5 for both datasets.

During evaluation, it is necessary to introduce the factor of the number of detected topics. Because it is impossible to pre-define the number of topics for this task. Therefore, we use two metrics to measure the performances: Top-10 F_1 versus number of detected topics (NDT) and accuracy versus false positive per topic (FPPT) [28]. Top-10 F_1 v.s. NDT measures the top-10 best detections, if we only need to measure top-n best results of a system without measuring false positives; while accuracy v.s. FPPT evaluates performances for each detected topics. That is, the former only evaluates the top-n best detections, while the latter measures the topic-wise performance. Moreover, accuracy (the truncated jaccard similarity) is more rigorous than top-10 F_1, since a low F_1 value does not model the coherence problem in topics [26], as discussed in [28].

For *top-10 F_1 v.s. NDT*, given a detected topic D_t, a ground truth topic G_t, the top-10 best F_1 scores are averaged to measure the performance

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

where Precision $= \frac{|D_t \cap G_t|}{|D_t|}$ is the precision, Recall $= \frac{|D_t \cap G_t|}{|G_t|}$ is the recall, and $|\cdot|$ denotes the number of webpages in a topic.

For accuracy v.s. FPPT, the accuracy is defined as

$$\text{Accuracy} = \frac{\#\text{Successful}}{\#\text{Groundtruth}}. \tag{16}$$

[2] [Online]. Available: www.nltk.org

A topic candidate D_t is recognized as a successful detection, if Normalized Intersected Ratio (NIR) $r = \frac{|D_t \cap G_t|}{|D_t \cup G_t|}$ is larger than a threshold [28]. In this paper, the threshold of NIR is set as 0.5. Note that if both methods have the same top-10 F_1 or accuracy score, the one with smaller NDT or FPPT achieves a better performance.

In all the experiments, SC is assigned the set of thresholds, $\{0.1, 0.5, 0.9\}$. In NMFR, the number of clusters is the set $\{100, 500, 900, 1300\}$, and the random walk parameter λ is 0.8.

B. Methods in Comparison Study

We specific experimental goal is to compare to the proposed approach with the state-of-the-art methods.

1) *Discriminative Probabilistic Models (DPM) [18]:* This baseline belongs to the text-modality based method, coming from the temporal discriminative probabilistic model for news streams. In the following experiments, we first resort to its offline version to embed documents into the discriminative feature space, and then the soft partition, vMF mixture model [3], is used to generate topics. DPM has reported better performance than LDA [6] in terms of discovering topics on several testbeds.

2) *Event-Clustering-Based Method (ECBM) [10]:* This baseline belongs to the multi-modality based method. Different from our scheme, the work [10] first clusters the tags in each time units, and then both the NDKs and the tag events are grouped into topics. Note that this approach involves many engineering details and hyper-parameters. We implement this method by ourself and report the best tuned results.

3) *Multi-Modality Graph (MMG) [42]:* The method belongs to the multi-modality method. Zhang *et al.* [42] the NDKs of videos and the text information, to build the similarity graph [29], and utilizes graph shift [23] on this graph to discover topics. Different from our method, this work assumes that the elements in a topic should be closely correlated. Therefore, MMG usually generates a small number of topics.

4) *Side-Information-Based Method (SIBM) [13]:* This method belongs to the text-modality based method. Chen *et al.* [13] first extract the hot searched queries from search engines, and refines the topics with an ad-hoc approach. This baseline demonstrates that our approach can achieve superior results without any supervised information on both MCG-WEBV and YKS.

5) *Maximal Cliques With Poisson Deconvolution (MCPD) [28]:* This baseline belongs to the text-modality based method. Different from our method, MCPD uses maximal cliques (MCs) as topic pattern, and further utilizes the PD approach to rank topics. The comparisons demonstrate the effectiveness of the way to exploit multi-modalities data in our method.

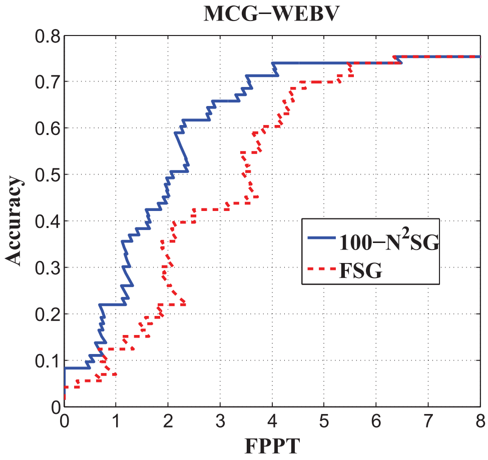


Fig. 4. Comparisons between k-N^{2}SG and FSG (best viewed in color).

C. Analysis of Our Approach

In this subsection, we use MCG-WEBV to test the effectiveness of our method:

1) Verify the effectiveness of k-N^2SG to partially reduce unstable correlations;
2) Compare the idea of learning a latent shared representation with the linear weight scheme (17);
3) Study the role of basis similarity to explain why the datum-wise fusion achieves better results than the baseline method (17);
4) Visualize the relationship between the μ_k and the size of C_k to make a better understanding of our method;
5) Discuss the scalability of the proposed method.

1) The Analysis of k-N^{2}SGs: In order to show that k-N^{2}SGs partially reduce the impact of noises, 100-N^{2}SGs and full similarity graphs (FSGs) are separatively built from HSGs, $TF{-}IDF{+}FV$ and $LDA{+}FV$. In this experiment, NMFR generates 4,240 and 3,445 topic candidates from the $TF{-}IDF{+}FV$ and the $LDA{+}FV$ 100-N^{2}HSGs respectively. Meanwhile, 4,245 and 3,289 topic candidates are generated by NMFR from the $TF{-}IDF{+}FV$ and $LDA{+}FV$ FSGs, respectively. Accuracy versus FPPT is used to evaluate the performances.

As illustrated in Fig. 4, 100-N^{2}HSG achieves a higher accuracy than that of FSG, when the FPPT value is smaller than 6. It indicates that k-N^{2}HSG indeed reduces a certain amount of unfavorable correlations from the less-constrained UGC, and thus increases the robustness of HSG. Also noticed that accuracies outputted by 100-N^{2}HSG increase faster than that of FSG, and as a result, the 100-N^{2}HSG outperforms FSG approximate 10% accuracy at FPPT $=1$ and about 25% one at FPPT $=3$, respectively. The increased performances of k-N^{2}HSG are consistent with our claim arguing against FSGs.

2) The Effectiveness of Latent Shared Representation: In the first experiment, to give fair comparisons, the PD approach [28] on the single HSG, $TF{-}IDF{+}FV$ or $LDA{+}FV$, is considered as the baseline method. Fig. 5 illustrates the effectiveness of exploiting multi-modalities data in topic detection.

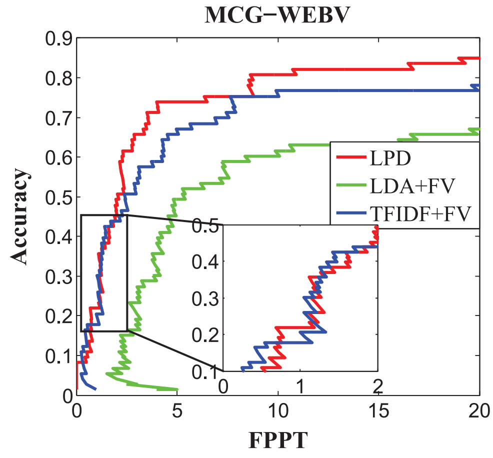


Fig. 5. PD on the single HSG versus LPD (best viewed in color).

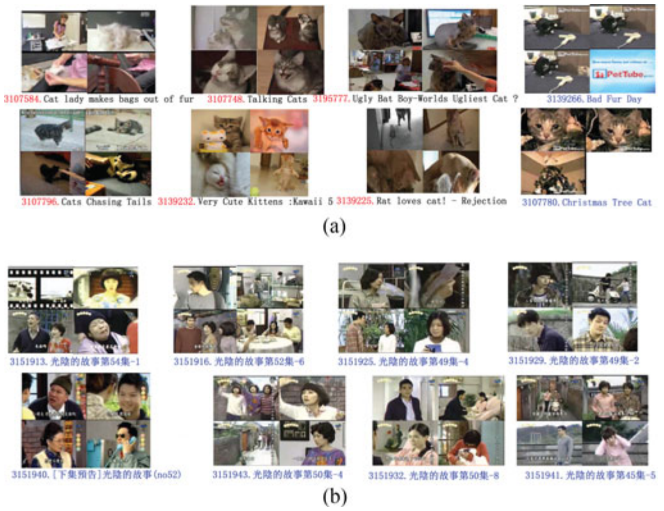

3107584. Cat lady makes bags out of fur
3107748. Talking Cats
3195777. Ugly Bat Boy-Worlds Ugliest Cat ?
3139266. Bad Fur Day
3107796. Cats Chasing Tails
3139232. Very Cute Kittens :Kawaii 5
3139225. Rat loves cat! - Rejection
3107780. Christmas Tree Cat
(a)
3151913. 光陰的故事第54集-1
3151916. 光陰的故事第52集-6
3151925. 光陰的故事第49集-4
3151929. 光陰的故事第49集-2
3151940. [下集預告]光陰的故事(no52)
3151943. 光陰的故事第50集-4
3151932. 光陰的故事第50集-8
3151941. 光陰的故事第45集-5
(b)

Fig. 6. Comparison between with and without LPD. False positive webpages are indicated in blue (best viewed in color). (a) Topic ranged by LPD is the interesting stories about cat. (b) Topic ranked by PD does not belong to the ground truth.

LPD largely outperforms the PD [28] method (denoted as the legend "LDA+FV") on the LDA+FV HSG. Compared with the results from the $TF-IDF$+FV HSG, LPD obtains very similar results when FPPT value is smaller than 2, but surpasses the PD method by about 5% accuracies where FPPT values are from 3 to 20. The consistently improved results indicate that the latent shared representation efficiently fuse the complementary information from the multiple HSGs.

Fig. 6 further visualizes the top-1 topics respectively ranked by LPD and PD on $TF-IDF$+FV for a vivid comparison, in which a webpage is represented by sampled keyframes and its title. Fig. 6 shows that LDP successfully detects one of the ground truth with NIR $= 0.62$. In contrast, although top-1 topic returned by PD in Fig. 6 is about TV series "Time Memory", this clustering does not belong to the ground truth. That is, it is not an interesting topic on social media. Moreover, the clustering about "Time Memory" is ranked as the 312th one by LPD, while the topic about cat is ranked as the 34th one by PD. The comparisons again indicate that: 1) web topic detection is not equal to the clustering task; and 2) the importance of exploiting multi-modal cues in topic detection.

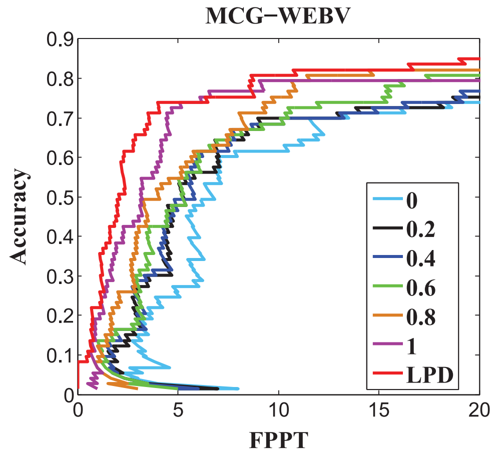


Fig. 7. Comparisons between the modality-level fusion and LPD (best viewed in color).

Next, we compared LPD with the conventional method, the modality-level fusion in (1), which assumes that every HSG has a non-negative weight

$$G_{\text{fusion}} = \alpha * G_{TF-IDF+FV} + (1-\alpha) * G_{LDA+FV} \quad (17)$$

where α is the non-negative weight. In our experiments, a set of α, $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, is used to tune the optimal fusion parameter. The resulting G_{fusion} is deconvoluted by the PD method [28] with topic candidates generated by NMFR. Note that the optimal hyperparameter α is determined in an exhaustive searching method [4]. By the grid search, we compared our method with the best results achieved the linear combination method in (17).

Fig. 7 illustrates the comparisons between our method and the modality-level fusion approach. As is unexpected, fusing the TF-IDF+FV and LDA+FV HSGs at the modality-level in (17) does not obtain the improved results. Interestingly, the LDA+FV HSG even drags down the performances of the TF-IDF+FV one, when α is assigned from 0.2 to 0.8. On the contrary, LPD consistently improves the performances on both HSGs. These results indicate that the datum-wise fusion is more suitable for UGC data than the modality-level approach.

3) The Role of Basis Similarity: As B^m serves as the basis to decompose the "background" similarities from the ones corrupted by noises, we wish to experimentally explain why sparse basis similarities help the datum-wise fusion. In this experiment, we begin with an analysis of the necessary of basis similarity and its sparsity, summarized in Fig. 8. Between the without basis similarity model "LPD-NoB^m" and the with basis similarity one "LPD-No$\|B^m\|_1$", it is clear that the basis similarity B^m plays an efficient role to increase the performances. For instance, when FPPT values range from 0 to 5, "LPD-No$\|B^m\|_1$" achieves better results than "LPD-NoB^m". In terms of the sparsity of the basis similarity, the performances of "LPD" largely outperform the counterpart "LPD-No$\|B^m\|_1$". These two sets

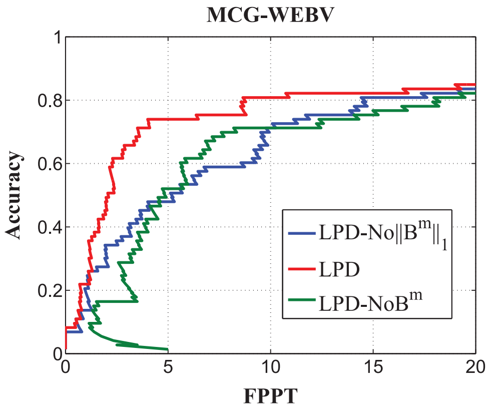


Fig. 8. Role of the sparse basis similarities (best viewed in color).

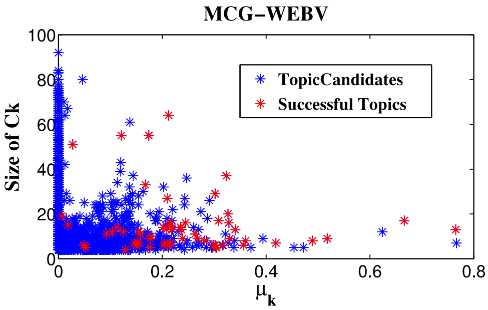


Fig. 9. Relationship between μ_k and $|C_k|$. Blue points are the topic candidates generated from HSGs, red ones are the successfully detected topics (best viewed in color).

of experiments indicate that the importance of imposing sparsity on the basis similarities B^m.

4) Analysis of the Relationship Between μ_k and $|C_k|$: Fig. 9 illustrates two important observations of LPD, in terms of the relationship between μ_k and $|C_k|$. The first is that about 85% μ_k $(k = 1, \ldots, K)$ are nearly equal to zeros, i.e., $\mu_k \leq 1e^{-5}$, although we do not explicitly enforce the spare constraint on μ_k. Moreover, the sizes of these zero-weight topics range from 4 to 90. This indicates that LPD is quit robust to the number of topic candidates, since only a few meaningful ones are selected. The second is that there is no close correlation between μ_k and $|C_k|$. That is, LPD does not favor the larger size of C_k by assigning a larger value to the corresponding μ_k. This observation also verifies the importance of the combination of the size of topic and its weight to identify a real topic.

5) Scalability: Since scalability is a major problem for tackling web data, we analysis the scalability of our method. Noticing that the scalability of a system involves the algorithms from the different components in a system, and their different implementations. The scalability of the proposed method involves three main components: k-N^{2}SG, candidates by NMFR, and LPD. The k-N^{2}SG can be efficiently approximated by recursive Lanczos bisection [12]. NMFR has been justified to have a good scalability ability [40]. LPD can be efficiently implemented by the scalable ADMM [45]. Therefore, the proposed method is theoretically scalable, if every important component is properly implemented.

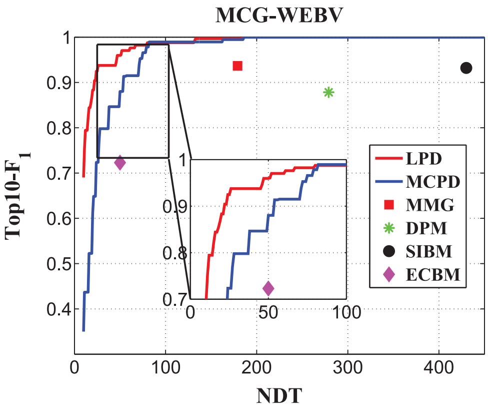


Fig. 10. Comparisons between the state-of-the-art methods and our method by Top-10 F_1 versus NDT on MCG-WEBV (best viewed in color).

D. Qualitative Comparisons with Other State-Of-The-Art Methods

In this subsection, we compare the proposed approach on other benchmark datasets. To make comparisons as meaningful as possible, we use the same experimental setups proposed by each dataset.

1) Web-Video Topic Detection in MCG-WEBV: Fig. 10 shows the comparison results by Top-10 F_1 versus NDT on MCG-WEBV. Our method achieves the highest Top-10 F_1 score than the others. Besides, Top-10 F_1 scores of our method increase quickly along with number of the generated topics. For instance, to achieve approximate 0.9 top-10 F_1 score, MCPD [28], MMG [42], SIBM [13] and DPM [18] generate 70, 179, 430, and 275 topics respectively on MCG-WEBV, while our method only generates 20 topic candidates.

The main explanation is that ECBM [10] totally depends on the clustering of tags, and then utilizes the visual and temporal consistency to link clusterings into topics. Naturally, a few noises in tags would greatly deteriorate the clustering results, due to the sparsity of tags per webpage, making its Top-10 F_1 remarkably low. Compared with SIBM [13], we can see that Top-10 F_1 is very close to MMBM [42]. Because the well selected key words from queries naturally filter out many false positives. On the other side, the NDT of SIBM [13] is much higher than MMBM [42], MCPD [28] and our approach. The explanation is that these key words from searching engines tend to have no correlation with these topics generated from social media. Among all these approaches, the PD based methods, both MCPD [28] and LPD, start with multi-granularity topics candidates, and then try to identify real topics in an unsupervised fashion. Therefore, Top-10 F_1 of the PD based approaches are much higher than that of all the other approaches.

To further evaluate the topic-wise performance, accuracy versus FPPT curves are plotted. As shown in Fig. 11, our approach is consistently better than MCPD [28], MMBM [42], DPM [18], SIBM [13] and ECBM [10]. Our system significantly outper-

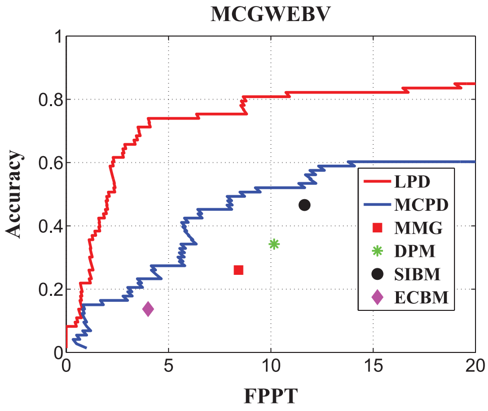


Fig. 11. Comparisons between the state-of-the-art methods and our method by accuracy versus FPPT on MCG-WEBV (best viewed in color).

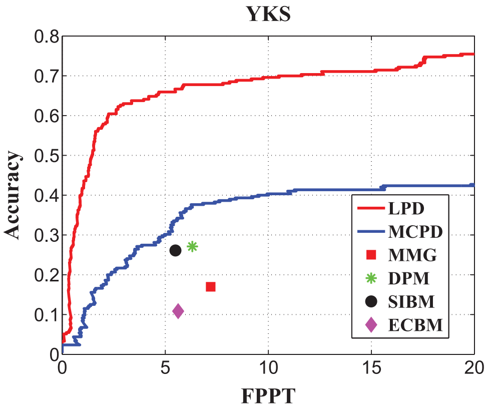


Fig. 13. Comparisons between the state-of-the-art methods and our method by accuracy versus FPPT on YKS (best viewed in color).

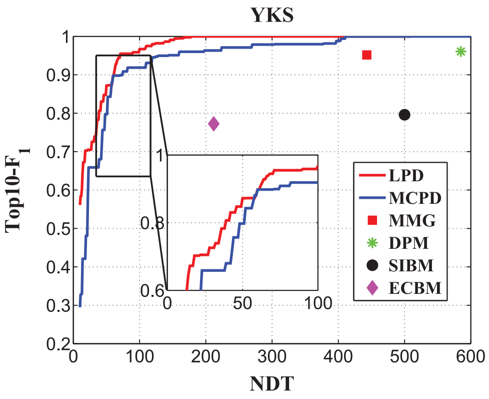


Fig. 12. Comparisons between the state-of-the-art methods and our method by Top-10 F_1 versus NDT on YKS (best viewed in color).

forms all the other state-of-the-art methods. For instance, when FPPT value equals to 5, the accuracy of our method is 0.78. As a contrast, the state-of-the-art MCPD [28] only obtains 0.28 accuracy. The explanation mainly includes two aspects: 1) the quality of topic candidates is critical to identify topics, as already observed in [28]; and 2) the datum-wise fusion is important to boost the performances, as discussed in Subsection V-C.

2) Web Topic Detection in YKS: YKS, a cross-platform dataset, requires to grasp more diverse types of topics than MCG-WEBV. Fig. 12 shows that our method consistently outperforms MCPD [28], MMBM [42], DPM [18], SIBM [13] and ECBM [10], if the same number of topics is generated. For instance, Top-10 F_1 of our method is 1, while ECBM [10] is 0.78, when 200 topics are generated for both methods. Compared with the results on MCG-WEBV dataset, Fig. 12 shows that both MMBM [42] and DPM [18] require to generate much more number of topics than that of our approach. For instance, MMBM [42] and DPM [18] have to generate 435 topic candidates and 590 ones in order to archive 0.95 Top-10 F_1, respectively. As a comparison, our approach only generates 100 topic candidates to obtain the same Top-10 F_1 score. This indicates the generalization ability of our approach across different data sets.

Fig. 13 further illustrates the accuracy versus FPPT curves on YKS. Our approach consistently outperforms these state-of-the-art methods. For instance, our method achieves an accuracy of 0.67, outperforming the MCPD [28] about the accuracy of 0.33. Moreover, if we compare Fig. 11 with Fig. 13, the accuracies of our method increase very fast. It means that our approach produces less false positives than the other state-of-the-art methods. Moreover, this observation again verifies the generalization ability of our method across different data sets.

Although the novelty of both top-k truncated similarity graph and candidates generated by NMFR is relative weak. Both contributions may seem small independently, as observed in Figs. 10–13, the resulting system improves the performances on both datasets significantly, e.g., from 0.31 accuracy for MCPD to 0.67 accuracy for our proposed system on YKS when FPPT $= 5$. This is a larger relative improvement in topic detection than that from the recent, state-of-the-art methods.

VI. Conclusion

In this paper, we have described a topic detection method by fusing multiple features into a latent shared graph, leading to the results surpassing the state-of-the-art methods in web topic detection. There are significant distinctions between the proposed LPD and the previous studies in exploiting the multimodal cues for topic detection.

1) We demonstrate the effectiveness of the datum-wise fusion for topic detection in exploiting the complementarity among the multiple modalities.
2) The proposed LPD enjoys both the advantage of the PD method [28] in achieving the high performances and that of MVL in exploiting complementary information among the multiple representations.
3) The datum-wise fusion scheme assumes no prior information about features, except the assumption on the sparsity of the basis similarity, in contrast to the weight method in (17), which assumes that there is nearly no noise in each feature representation.

The promising results of this paper motivate a further examining of the LPD-based topic detection. First, more effective constraints about the basis similarity, like low-rank, may bring more interesting merits over the sparsity used here. Moreover, online optimization of LPD scales up well to large-scale problems [8].

References

[1] M. Aiello *et al.*, "Sensing trending topics in twitters," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1268–1282, Oct. 2013.

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," in *Proc. DARPA Broadcast News Transcription Understanding Workshop*, Feb. 1998, pp. 194–218.

[3] A. Banerjee and S. Basuy, "Topic models over text streams: a study of batch and online unsupervised learning," in *Proc. SIAM Int. Conf. Data Mining*, pp. 431–436, 2007.

[4] J. Bergstra and B. Yoshua, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.

[5] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. Int. Conf. Data Mining*, 2004, pp. 19–26.

[6] D. Blei, M. David, A. Ng, M. Jordan, and J. Lafferty, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[7] D. Blei and J. Lafferty, "A correlated topic model of science," *Annal. Appl. Sci.*, vol. 1, pp. 17–35, 2007.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[9] T. Brants, F. Chen, and A. Farahat, "A system for new event detection," in *Proc. ACM SIGIR Conf. Res. Develop. Inform. Retrieval*, 2003, pp. 330–337.

[10] J. Cao, C. Ngo, Y. Zhang, and J. Li, "Tracking web video topics: Discovery, visualization, and monitoring," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1835–1846, Dec. 2011.

[11] K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan,"Multi-view clustering via canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 129–136.

[12] J. Chen, H. Fang, and Y. Saad, "Fast approximate $k$nn graph construction for high dimensional data via recursive Lanczos bisection," *J. Mach. Learn. Res.*, vol. 10, pp. 1989–2012, 2009.

[13] T. Chen, C. Liu, and Q. Huang, "An effective multi-clue fusion approach for web video topic detection," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 781–784.

[14] S. Cheung and A. Zakhor, "Efficient video similarity measurement with video signature," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 59–74, Jan. 2003.

[15] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[16] B. Han, P. Cook, and T. Baldwin, "Automatically constructing a normalisation for microblogs," in *Proc. Joint Conf. Empirical Meth. Natural Languages Process. Comput. Natural Language Learn.*, 2012, pp. 421–432.

[17] Q. He, K. Chang, and E. Lim, "Analyzing feature trajectories for event detection," in *Proc. ACM SIGIR Res. Develop. Inform. Retrieval*, 2007, pp. 207–214.

[18] Q. He, K. Chang, E. Lim, and A. Banerjee, "Keep it simple with time: A re-examination of probabilitic topic detection models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1795–1808, Oct. 2010.

[19] T. Hofmann, "Probabistical latent semantic indexing," in *Proc. ACM SIGIR Conf. Res. Develop. Inform. Retrieval*, 1999, pp. 50–57.

[20] J. Kludas, E. Bruno, and S. Marchand-Mailet, "A general model for multiple view unsupervised learning," in *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics*. New York, NY, USA: Springer, 2008, pp. 147–159.

[21] A. Kumar, P. Rai, and H. Daumé, "Co-regularized multi-view spectral clustering," in *Proc. Neural Inf. Process. Syst.*, 2011, pp. 1412–1421.

[22] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multipler method for exact recovery of corrupted low-rank matrices," *CoRR*, 2013. [Online]. Available: http://arxiv.org/abs/1000.5055

[23] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link lda: Joint models of topic and author community," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 338–349.

[24] Z. -Q. Luo, "On the linear convergence of the alternating direction method of multipliers," *CoRR*, 2012. [Online]. Available: http://arxiv.org/abs/1208.3922

[25] S. Neo, Y. Ran, H. Goh, Y. Zheng, T. Chua, and J. Li, "The use of topic evaluation to help users browse and find answer in new video corpus," in *Proc. Int. Conf. ACM Multimedia*, 2007, pp. 198–207.

[26] D. Newman, E. Bonilla, and W. Buntine, "Improving topic coherence with regularized topic models," in *Neural Inform. Process. Syst.*, 2011, pp. 496–504.

[27] Y. Pan, H. Lai, C. Liu, Y. Tang, and S. Yan, "Rank aggregation via low-rank and structured-sparse decomposition," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 3021–3028.

[28] J. Pang *et al.*, "Unsupervised web topic detection using a ranked clustering-like pattern across similarity cascades," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 843–853, Jun. 2015.

[29] S. Papadopoulous, C. Zigkolis, Y. Kompatsiaris, and A. Vakali, "Cluster-based landmark and event detection on tagged photo collections," *IEEE Multimedia*, vol. 18, no. 1, pp. 52–63, Jan. 2011.

[30] D. Putthividhy, H. Attias, and S. Magarajan, "Topic regression multimodal latent dirichlet allocation for image annotation," in *Proc. IEEE Proc. Comput. Vis. Pattern Recog.*, vol. 1, 2010, pp. 3408–3415.

[31] J. Sánchez, T. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[32] D. Shahaf and C. Guestrin, "Connecting the dots between news articles," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 623–632.

[33] A. Sun and M. Hu, "Query-guided event detection from news and blog streams," *IEEE Trans. Syst., Man Cybern. A*, vol. 41, no. 5, pp. 834–839, Sep. 2011.

[34] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical dirichlet processes," *J. Amer. Statistical Assoc.*, vol. 101, pp. 1566–1581, 2006.

[35] C. Wang, M. Zhang, S. Ma, and L. Ru, "Automatic online news issue construction in web environment," in *Proc. Int. Conf. World Wild Web*, 2008, pp. 457–466.

[36] X. Wu, G. Hauptmann, and C. Ngo, "Novelty detection for crosslingual news story with visual duplicates and speech transcripts," in *Proc. Int. Conf. ACM Multimedia*, 2007, pp. 168–177.

[37] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *CoRR*, 2013. [Online]. Available: http://arxiv.org/abs/1304.5634

[38] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. ACM SIGIR Res. Develop. Inf. Retrieval*, 2003, pp. 267–273.

[39] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *Proc. ACM SIGIR Res. Develop. Inf. Retrieval*, 1998, pp. 28–36.

[40] Z. Yang, T. Hao, O. Dikmen, X. Chen, and E. Oja, "Clustering by nonnegative matrix factorization using graph random walk," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1079–1087.

[41] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang, "Robust late fusion with rank minimization," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 3021–3028.

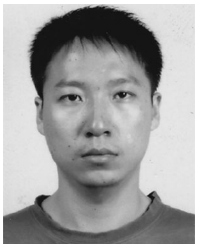
[42] Y. Zhang *et al.*, "Cross-media topic detection: a multi-modality fusion framework," in *Proc. Int. Conf. Multimedia Expo*, 2013, pp. 1–6.

[43] W. X. Zhao *et al.*, "Comparing twitter and traditional media using topic models," in *Proc. Eur. Conf. Adv. Inf. Retrieval*, 2011, pp. 338–349.

[44] D. Zhou and C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 1159–1166.

[45] L. Zhou and J. Kwok, "Fast stochastic alternating direction method of multipliers," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1–9.

[46] A. Zien and C. Ong, "Multiclass multiple kernel learning," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 1191–1198.



Junbiao Pang received the B.S. and M.S. degrees in computational fluid dynamics and computer science from the Harbin Institute of Technology, Harbin, China, in 2002 and 2004, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011.

He was an Assistant Professor of computer science and technology with the Beijing University of Technology (BJUT), Beijing, China, from 2011 to 2013. He is currently an Associate Professor with the College of Metropolitan Transportation, BJUT. He has authored or coauthored more than 20 academic papers in publications such as the IEEE Transactions on Image Processing, ECCV, ICCV, and ACM Multimedia. His research interests include multimedia and machine learning.

Fei Tao received the B.E. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2014, and he is currently working toward the M.S. degree in computer science and technology at the University of the Chinese Academy of Sciences, Beijing, China.

His research interests include machine learning, image content analysis, and information retrieval.

Chunjie Zhang received the B.E. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2006, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012.

He is currently a Faculty Member with the University of the Chinese Academy of Sciences, Beijing, China. His research interests include machine learning, image content analysis, and object categorization.

Weigang Zhang received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2003, 2005, and 2016, respectively.

He is an Associate Professor with the School of Computer Science and Technology, HIT, Weihai, China, and also a Postdoctoral Researcher with the University of Chinese Academy of Sciences, Beijing, China. His research interests include multimedia computing and computer vision.

Qingming Huang (M'04–SM'08) received the B.S. degree in computer science and Ph.D. degree in computer engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is currently a Professor with the University of the Chinese Academy of Sciences (CAS), Beijing, China, and an Adjunct Professor with the Institute of Computing Technology, CAS, China, and with the Beijing University of Technology, Beijing, China. He has authored or coauthored more than 300 academic papers in prestigious international journals including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON IMAGE PROCESSING, and top-level conferences such as ACM Multimedia, ICCV, CVPR, AAAI, IJCAI and VLDB. His research interests include multimedia computing, image processing, computer vision, pattern recognition, and machine learning.

Dr. Huang is an Associate Editor of *Acta Automatica Sinica* and a Reviewer of various international journals including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He has served as Program Chair, Track Chair, Area Chair, and TPC Member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, and PSIVT.

Baocai Yin (M'15) received the M.S. and Ph.D. degrees in computational mathematics from the Dalian University of Technology, Dalian, China, in 1988 and 1993, respectively.

He is currently a Professor with the Department of Electronic Information and Electrical Engineering, Dalian University of Technology. He is also the Director of Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing, China. He has authored or coauthored more than 200 academic papers in prestigious international journals including the IEEE TRANSACTIONS ON MULTIMEDIA and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and top-level conferences such as INFOCOM and ACM SIGGRAPH. His research areas include multimedia, image processing, computer vision, and pattern recognition.

Dr. Yin is currently an Editorial Member for the *Journal of Information and Computational Science* (USA).