Learning Manifold Representation from Multimodal Data for Event Detection in Flickr-Like Social Media

Zhenguo Yang[1(✉)], Qing Li[1,2], Wenyin Liu[2], and Yun Ma[1]

[1] Department of Computer Science, City University of Hong Kong, Hong Kong, China
yzgcityu@gmail.com, itqli@cityu.edu.hk, mayun371@gmail.com
[2] Multimedia-software Engineering Research Center, City University of Hong Kong, Hong Kong, China
liuwenyin@gmail.com

Abstract. In this work, a three-stage social event detection model is devised to discover events in Flickr data. As the features possessed by the data are typically heterogeneous, a multimodal fusion model (M^2F) exploits a soft-voting strategy and a reinforcing model is devised to learn fused features in the first stage. Furthermore, a Laplacian non-negative matrix factorization (LNMF) model is exploited to extract compact manifold representation. Particularly, a Laplacian regularization term constructed on the multimodal features is introduced to keep the geometry structure of the data. Finally, clustering algorithms can be applied seamlessly in order to detect event clusters. Extensive experiments conducted on the real-world dataset reveal the M^2F-LNMF-based approaches outperform the baselines.

Keywords: Social media analytics · Multimedia content analysis · Multimodal fusion · Manifold learning · Event detection

1 Introduction

The popularity of Flickr-like photo-sharing social media services has resulted in huge amounts of user-contributed images available online, attracting researchers to link the data to numerous real-world concepts. Social event detection (SED) from Flickr data aims to discover the real-world events that are attended by people and can be represented by user-contributed multimedia data. Instances of such events could include concerts, public celebrations, annual conventions, local gatherings, sports events, etc.

SED from Flickr data is deemed to be more challenging than the event detection tasks in textural data [4,11] as the data is heterogeneous. For instance, the Flickr images possess context features including time-taken, user identity, location, tags and visual content, etc. Such features will be helpful for capturing the similarities among the social media documents and, in turn, for identifying

H. Gao et al. (Eds.): DASFAA 2016 Workshops, LNCS 9645, pp. 160–167, 2016.
DOI: 10.1007/978-3-319-32055-7_14

event clusters and their associated documents. However, the heterogeneous features are hard to be exploited by traditional clustering or classification models seamlessly. To address the problem, early fusion and late fusion are widely-used strategies. Late fusion is expensive in terms of learning effort and the result may not be good since each modality might be poor. As a result, early fusion strategy is more popular. However, early fusion models usually construct multiple affinity graphs [1,2] with intensive computations, making them not adaptive in dealing with social media data due to its large quantity and high updating rate.

In this work, a social event detection model is designed to discover events from photo-sharing social media sites, which consists of three stages. In the first stage, we propose a multimodal fusion (M^2F) model to learn fused features from the heterogeneous feature modalities. Particularly, M^2F exploits a unimodal soft-voting strategy to learn comparable and robust vote features, which expresses the data samples by their neighborhood information. Furthermore, M^2F exploits a reinforcing model to learn the vote propagations among the multimodal features and achieve fused features. In the second stage, we exploit a Laplacian non-negative matrix factorization model, denoted as LNMF, to extract compact manifold representation from the fused features. In particular, the Laplacian regularization term is constructed based on the multimodal features, which tends to learn similar manifold representation for the samples that are close in the fused feature space. In the third stage, clustering algorithms can be applied on the manifold representation learned by M^2F-LNMF to discover event clusters. Particularly, incorporating density knowledge or label information in the initialization of the center-based clustering algorithm will give significant improvement on the performance.

The rest of the paper is organized as follows. In Sect. 2, the related work is reviewed. In Sect. 3, the proposed three-stage social event detection model is presented. In Sect. 4, extensive experiments are conducted and analyzed. Finally, we offer conclusions in Sect. 5.

2 Related Work

2.1 Social Event Detection in Multimedia Social Media Streams

To detect events from Flickr data, researchers usually employed classification models. Liu et al. [7] trained various models like KNN, SVM, decision tree and random forest to classify the images into the events they depict. Chen et al. [5] developed a system to discover semantic concept in videos by exploiting Web images and their associated tags, and trained a SVM model for predictions. However, most of them are designed for domain-specific events that are well-defined in advance, making them not adaptive in dealing with the variety of events in social media. Nitta et al. [8] constructed similarity graphs and applied community detection to identify subgraphs that could be landmarks or events. However, these approaches require pair-wise similarity calculations, which are computation-intensive and memory-consuming.

2.2 Multimodal Feature Fusion

A number of early fusion models have been proposed. For instance, Cai et al. [2] computed a multimodal Laplacian matrix by integrating the individual affinity matrix on each modality, and further learned a low-dimensional feature space by introducing a penalty for each modality. Julien et al. [1] proposed a unifying graph-based framework, which combines both visual and textual information. Petkos et al. [9] trained a classifier to acquire an indicator matrix showing whether two images could be in the same event. Furthermore, spectral clustering was performed on the indicator matrix to discover event clusters. However, these approaches have high computational complexity in constructing multiple affinity matrices, which makes them hardly to be applied in dealing with large-scale of social media data.

3 Multimodal Fusion and Manifold Learning for SED

In this section, the proposed three-stage SED model is presented. Particularly, we preprocess the data by introducing the multimodal similarity metrics before the three-stage event detection process, and present each stage in each subsection.

3.1 Similarity Metrics for the Feature Modalities

Considering Flickr images possess multiple context features, such as *Time*, *Location*, *Tags*, *User identity*, etc., we define the similarity metrics for them as a preprocessing step. Specifically, *Time* similarity is measured by using an inverse function [13] on the time interval. *Tags* similarity is calculated on the associated tags by using *Jaccard index*. *User* similarity is defined as a binary indicator to show whether the images were taken by the same user. *Location* similarity can be calculated by using *Haversine* formula on the *latitude* and *longitude* attributes. Particularly, we adopt perceptual hashes ($pHash$) to evaluate the visual similarity as it is advantageous in terms of efficiency and memory cost.

3.2 Multimodal Fusion Model ($\mathbf{M^2F}$)

In order to exploit the rich context features associated with social media data, we propose a multimodal fusion ($\mathrm{M^2F}$) model, which consists of unimodal feature voting step and vote feature reinforcing step. In the first step, $\mathrm{M^2F}$ collects the votes from neighboring dictionary images for a given image, which can be obtained by computing their similarities on each feature modality based on the defined metrics. The image dictionary can be obtained from the images with labels. For a number of M feature modalities, d patterns in the image dictionary, and n samples in the image collection, vote matrices for the modalities, denoted as $F^1 \in \mathbb{R}^{d\times n}, ..., F^M \in \mathbb{R}^{d\times n}$ respectively, can be obtained based on the unimodal feature voting processes.

Furthermore, M^{2}F exploits a reinforcing model to learn the vote propagations among the feature modalities and achieve fused features in the second step, as specified in Eq. (1),

$$\begin{cases} (F^1)^{(n)} = p_1(F^1)^{(0)} + (1-p_1)(F^2)^{(n-1)} \\ (F^2)^{(n)} = p_2(F^2)^{(0)} + (1-p_2)(F^3)^{(n-1)} \\ \vdots \\ (F^m)^{(n)} = p_m(F^m)^{(0)} + (1-p_m)(F^1)^{(n-1)} \end{cases} \tag{1}$$

where $(F^1)^{(n)}$, $(F^2)^{(n)}$,...,$(F^m)^{(n)}$ indicate the results at n-th iteration, $n = 0$ denotes the original affinity matrices, and p_m is the parameter for the m-th modality. The iterative process will be terminated until $(F^1)^{(n)}$ is convergent to $(F^1)^{(n-1)}$. As a result, the fused features $X \in \mathbb{R}^{M\times N}$ can be assigned as the converged result of $(F^1)^{(n)}$.

3.3 LNMF-Based Manifold Learning

Based on the fused features achieved by M^{2}F, manifold learning models such as NMF, PCA, ICA, etc., can be applicable to deal with multimodal tasks. Considering the non-negativity of the fused features achieve by M^{2}F, we introduce graph regularized non-negative matrix factorization model [2] by defining a Laplacian term based on the fused features, which is denoted by LNMF, to extract compact representation. Formally, given the fused features learned by M^{2}F in matrix form, denoted as $X \in \mathbb{R}^{d\times n}$, manifold learning aims to learn k-dimensional ($k < d$) hidden data representation, denoted as $H \in \mathbb{R}^{k\times n}$. by approximating the original data matrix with two non-negative matrices $W \in \mathbb{R}^{d\times k}$ and $H \in \mathbb{R}^{k\times n}$. Particularly, a Laplacian term constructed from the fused features is introduced to keep the geometry structure of the data in the manifold learning process.

(1) Objective Function. The objective function of LNMF is defined as follows,

$$\begin{array}{ll} \underset{W,H}{\text{minimize}} & \|X - WH\|^2 + \frac{\lambda}{2}\sum_{i,j}^{n} A_{i,j}\|h_i - h_j\|^2 \\ \text{subject to} & W \geq 0, H \geq 0. \end{array} \tag{2}$$

where the first term aims to minimize the reconstruction errors and the second one is the graph regularization term. $A_{i,j}$ is the affinity between image i and j, which can be calculated as follows,

$$A_{i,j} = \frac{\sum_{f=1}^{d} F_{f,i}F_{f,j}}{\sqrt{\sum_{f=1}^{d}(F_{f,i})^2}\sqrt{\sum_{f=1}^{d}(F_{f,j})^2}} \tag{3}$$

where $F \in \mathbb{R}^{d\times n}$ is the element in the fused features. Particularly, the graph regularization term can be represented by a Laplacian term as follows,

$$\begin{aligned} &\frac{1}{2}\sum_{i,j}^{n} A_{i,j}\|H_i - H_j\|^2 = \sum_{i=1}^{n} D_{i,i}H_iH_i^T - \sum_{i,j}^{n} A_{i,j}H_iH_j^T \\ &= Tr(HDH^T) - Tr(HAH^T) = Tr(HLH^T) \end{aligned} \tag{4}$$

where $Tr(\cdot)$ denotes the trace of matrix. Furthermore, we arrive at the following formulation,

$$\begin{array}{ll} \underset{W,H}{\text{minimize}} & \|X - WH\|^2 + \lambda Tr(HLH^T) \\ \text{subject to} & W \geq 0, H \geq 0. \end{array} \tag{5}$$

(2) Optimizations. We denote the objective function of LNMF as $J(W, H)$, which can be minimized in a gradient descent manner by adopting additive updating rules as follows.

$$\begin{cases} (W^{(n+1)})_{i,j} = (W^{(n)})_{i,j} - \gamma(\nabla_{W^{(n)}} J(W, H))_{i,j} \\ (H^{(n+1)})_{i,j} = (H^{(n)})_{i,j} - \gamma(\nabla_{H^{(n)}} J(W, H))_{i,j} \end{cases} \tag{6}$$

where the indicator (n) denotes the n-th iteration, γ is the step size parameter controlling the learning rate, $\nabla_{W^{(n)}} J(W, H)$ and $\nabla_{H^{(n)}} J(W, H)$ denotes the partial derivatives of $J(W, H)$ with respect to $W^{(n)}$ and $H^{(n)}$ respectively. Updating W and H will be terminated until converged.

$$\begin{cases} \nabla_{W^{(n)}} J(W, H) = -2XH^T + 2W^{(n)}HH^T \\ \nabla_{H^{(n)}} J(W, H) = -2W^T X + 2W^T WH^{(n)} + 2\lambda H^{(n)} L \end{cases} \tag{7}$$

3.4 Event Clustering

To detect event clusters in the image collections, clustering algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and K-Means, can be applied seamlessly on the manifold feature representation learned by M^{2}F-LNMF. Considering the randomness of the initial centers of K-Means, we initialize the cluster centers by exploiting unsupervised density knowledge or the semi-supervised label information, denoted by DKM and SKM respectively. For simplicity, we denote the social event detection algorithms applying DBSCAN, K-Means, DKM and SKM in the third stage as M^{2}F-LNMF-DBSCAN (MLD), M^{2}F-LNMF-K-Means (MLK), M^{2}F-LNMF-DKM (MLDK) and M^{2}F-LNMF-SKM (MLS), respectively.

4 Experiments

In this section, we conduct a series of experiments on MediaEval Social Event Detection 2014 [10], which contains 110,541 Flickr images that are related to 6,635 events in total. The images in the dataset have been associated with some context features, such as image identifier, geo-tags, time-stamp, user identifier and tags, etc. However, 80 % of the geo-tags are not available.

4.1 Baseline Algorithms

The baselines include a number of recently proposed methods on SED and multimodal fusion tasks, such as graph-based Multimodal Spectral Clustering

(MMSC) [2], SVD-based Multimodal Clustering (SVD-MC) [13], Semi-supervised Multimodal Clustering (SMC) [14], Constrained Incremental Clustering via Ranking (CICR) [12]. In addition to LNMF, GRBM [6], PCA, ICA are applicable in the manifold learning stage, and we implement them for comparisons and denote them as M^{2}F-LNMF (ML), M^{2}F-GRBM (MG), M^{2}F-PCA (MP), M^{2}F-ICA (MI), respectively. All the methods are tested on MediaEval SED 2014 tasks via the Normalized Mutual Information (NMI), which is a standard technique for evaluating the quality of clusters. The value of NMI is between 0 and 1, and a larger value is preferred.

4.2 Experimental Results

The experimental results for the SED tasks are shown in Table 1, from which some interesting observations can be concluded. Firstly, the proposed M^{2}F-LNMF-based approaches, i.e., MLK, MLD, MLDK and MLS, outperform the baselines, giving significant improvement on the performance. Secondly, compared to SMC, which has no manifold learning process, the performance is improved by 7 % at most, indicating the effectiveness of the LNMF model. Thirdly, MLD and MLS exploit density knowledge and labeled data to initialize the centers, outperforming the approaches using either of them. Note that the superscript "*" denotes the result outperforms the best one from the baselines.

Table 1. NMI achieved by the algorithms

	MMSC	SVD-MC	CICR	SMC	SMR	JMSR	MLD	MLK	MLDK	MLS
NMI	0.5982	0.8940	0.9024	0.9113	0.9413	0.9417	0.9475*	0.9426*	0.9536*	**0.9751***

4.3 Evaluation on Manifold Learning and Event Clustering Models

In addition to LNMF, we implement PCA, ICA, GRBM [6] in the second stage of the SED model for comparisons. The experimental results are shown in Table 2, which indicates LNMF outperforms the other models on the SED tasks. On the other hand, DKM and SKM that have incorporated density knowledge or label information achieve better performance than using either DBSCAN or K-Means merely.

Table 2. NMI achieved by the combinations of the manifold learning and clustering models

	DBSCAN	K-Means	DKM	SKM
M^{2}F-GRBM (MG)	0.8127	0.8198	0.8345	0.8366
M^{2}F-ICA (MI)	0.9275	0.9164	0.9487*	0.9589*
M^{2}F-PCA (MP)	0.9365	0.9407	0.9512*	0.9678*
M^{2}F-LNMF (ML)	**0.9475***	**0.9426***	**0.9536***	**0.9751***

4.4 Evaluation on the Dictionary Scale

In the process of M^{2}F, an image dictionary is used for the unimodal feature voting, and we evaluate the impact of the dictionary scale as shown in Fig. 1, where MPS denotes MP-SKM, and so on. From the figure some interesting observations can be revealed. Firstly, the NMI values achieved by MLS and MPS increase and tend to be sable with the scale of dictionary increases. The reason can be explained by that a larger dictionary will capture more patterns of the data. However, a too large dictionary could not capture more than the actual number of patterns in the data collection, i.e., the dictionary must contain replicated patterns. As a result, As a result, LNMF and PCA can deal with the replicated pattern problem well, while ICA is negatively impacted as shown in the figure. On the other hand, a dictionary that is too small may not be complete in expressing the data collections. Secondly, the proposed MLS is more effective for the current SED tasks outperforming the baselines.

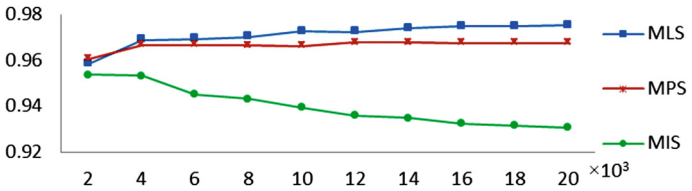

Fig. 1. Impact of the dictionary scale

5 Conclusion

In this paper, we propose a three-stage SED model, i.e., M^{2}F-based multimodal fusion, LNMF-based manifold learning and event clustering. Firstly, fused features integrating the multimodal features are achieved by M^{2}F. Furthermore, compact manifold representation is learned by LNMF, keeping the geometry structure of the data in the learning process. Finally, clustering algorithms can be applied on the manifold learned by M^{2}F-LNMF seamlessly to discover event clusters. Particularly, the hybrid clustering algorithm gives significant improvement on the performance. The experiments conducted on the real-world dataset manifest the effectiveness of the M^{2}F-LNMF based event detection approaches.

Acknowledgments. We would like to thank Dr. Zheng Lu, Mr. Min Cheng and Mr. Yangbin Chen for the discussions.

References

1. Ah-Pine, J., Csurka, G., Clinchant, S.: Semi-supervised visual and textual information fusion in CBMIR using graph-based methods. ACM Trans. Inf. Syst. (TOIS) **33**(2), 9 (2015)

2. Cai, X., Nie, F., Huang, H., Kamangar, F.: Heterogeneous image feature integration via multi-modal spectral clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1977–1984. IEEE Press (2011)
3. Cai, D., He, X., Han, J., Huang, T.S.: Graph regularized nonnegative matrix factorization for data representation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1548–1560 (2011)
4. Cai, Y., Li, Q., Xie, H., Wang, T., Min, H.: Event relationship analysis for temporal event search. In: Meng, W., Feng, L., Bressan, S., Winiwarter, W., Song, W. (eds.) DASFAA 2013, Part II. LNCS, vol. 7826, pp. 179–193. Springer, Heidelberg (2013)
5. Chen, J., Cui, Y., Ye, G., Liu, D., Chang, S.F.: Event-driven semantic concept discovery by exploiting weakly tagged internet images. In: International Conference on Multimedia Retrieval, pp. 1–8. ACM (2014)
6. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
7. Liu, X., Huet, B.: Heterogeneous features and model selection for event-based media classification. In: 3rd ACM International Conference on Multimedia Retrieval, pp. 151–158. ACM (2013)
8. Nitta, N., Kumihashi, Y., Kato, T., Babaguchi, N.: Real-world event detection using flickr images. In: Gurrin, C., Hopfgartner, F., Hurst, W., Johansen, H., Lee, H., O'Connor, N. (eds.) MMM 2014, Part II. LNCS, vol. 8326, pp. 307–314. Springer, Heidelberg (2014)
9. Petkos, G., Papadopoulos, S., Kompatsiaris, Y.: Social event detection using multimodal clustering and integrating supervisory signals. In: 2nd ACM International Conference on Multimedia Retrieval, p. 23. ACM (2012)
10. Petkos, G., Papadopoulos, S., Mezaris, V., Kompatsiaris, Y.: Social event detection at MediaEval 2014: Challenges, datasets, and evaluation. In: MediaEval 2014 Workshop (2014)
11. Rao, Y., Li, Q.: Term weighting schemes for emerging event detection. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, pp. 105–112 (2012)
12. Sutanto, T., Nayak, R.: Ranking based clustering for social event detection. In: MediaEval 2014 Workshop, 1263, pp. 1–2 (2014)
13. Yang, Z., Li, Q., Lu, Z., Ma, Y., Gong, Z., Pan, H.: Semi-supervised multimodal clustering algorithm integrating label signals for social event detection. In: IEEE International Conference on Multimedia Big Data, pp. 32–39. IEEE (2015)
14. Yang, Z., Li, Q., Lu, Z., Ma, Y., Gong, Z., Pan, H., Chen, Y.: Semi-supervised multimodal fusion model for social event detection on web image collections. Int. J. Multimed. Data Eng. Manage. **6**(4), 1–22 (2015)