Cross-Domain Collaborative Learning in Social Multimedia

Shengsheng Qian[1], Tianzhu Zhang[1], Richang Hong[2], Changsheng Xu[1]
[1]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
[2]School of Computer and Information, Hefei University of Technology
{shengsheng.qian, tzzhang, csxu}@nlpr.ia.ac.cn, hongrc@hfut.edu.cn

ABSTRACT

Cross-domain data analysis is one of the most important tasks in social multimedia. It has a wide range of real-world applications, including cross-platform event analysis, cross-domain multi-event tracking, cross-domain video recommendation, etc. It is also very challenging because the data have multi-modal and multi-domain properties, and there are no explicit correlations to link different domains. To deal with these issues, we propose a generic Cross-Domain Collaborative Learning (CDCL) framework based on non-parametric Bayesian dictionary learning model for cross-domain data analysis. In the proposed CDCL model, it can make use of the shared domain priors and modality priors to collaboratively learn the data's representations by considering the domain discrepancy and the multi-modal property. As a result, our CDCL model can effectively explore the virtues of different information sources to complement and enhance each other for cross-domain data analysis. To evaluate the proposed model, we apply it for two different applications: cross-platform event recognition and cross-network video recommendation. The extensive experimental evaluations well demonstrate the effectiveness of the proposed algorithm for cross-domain data analysis.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services

Keywords

social media, cross-domain collaborative learning, multi-modality

1. INTRODUCTION

With the rapid development of Internet, there are more and more social media sites (e.g., Flickr, YouTube, Facebook, and Google News), which make people be able to conveniently generate and share rich social multimedia content online, including multimedia documents, social links. As a result, a popular event topic that is happening around us and around the world can spread very fast in different media sites, and there are substantial amounts of media data with multi-modality (e.g., images, videos, and text). For example, as shown in Figure 1(a), a hot topic (e.g., United States Presidential Election) emerges, there are many relevant documents (images and text) in different platforms, such as Goodle News, Flickr. These documents have different perspectives, official on Google News, and personal comments and interesting photos on Flickr. If we aggregate the relevant data across different platforms, they can complement and enhance each other, especially when the strengths of one domain complement the weaknesses of the other. For example, a lot of interesting comments and photos by Web users on Flickr can complement few official reports on Google News. Moreover, the images of official reports generally captured by journalists can focus on the targets of a specific event perfectly on Google News, while most uploaded images which are typically captured by users are not professional on Flickr. To better understand what happens across multiple platforms, it is better to make use of the virtues of different information sources via collaborative learning algorithm. In social multimedia, the cross-domain collaborative learning (CDCL) is an important task for knowledge mining as it aims at discovering collective and subjective information, which may be more beneficial to users than a single domain in many applications such as cross-platform event analysis [1], cross-domain multi-event tracking [2], cross-domain collaboration recommendation [3].

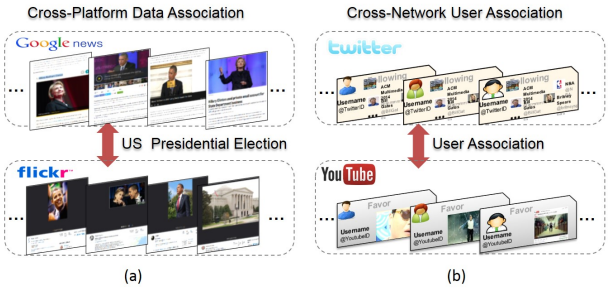


Figure 1: Two different cross-domain scenarios: (a) cross-platform data association (b) cross-network user association(All photos via Flickr under Creative Commons License).

In real-world scenarios, different domains can bridge the domain gap via the shared domain information, such as event topics, social links. As shown in Figure 1, we illustrate a toy example and show two different cross-domain scenarios including cross-platform data association and cross-network user association. In Figure 1(a), we show an example about the cross-platform data association. Here, Google News and Flickr are regarded as two domains, and their documents are associated via the shared event topic "United States Presidential Election". In Figure 1(b), it shows an example about the cross-network user association. In the two different network-

MM'15, October 26-30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806234.

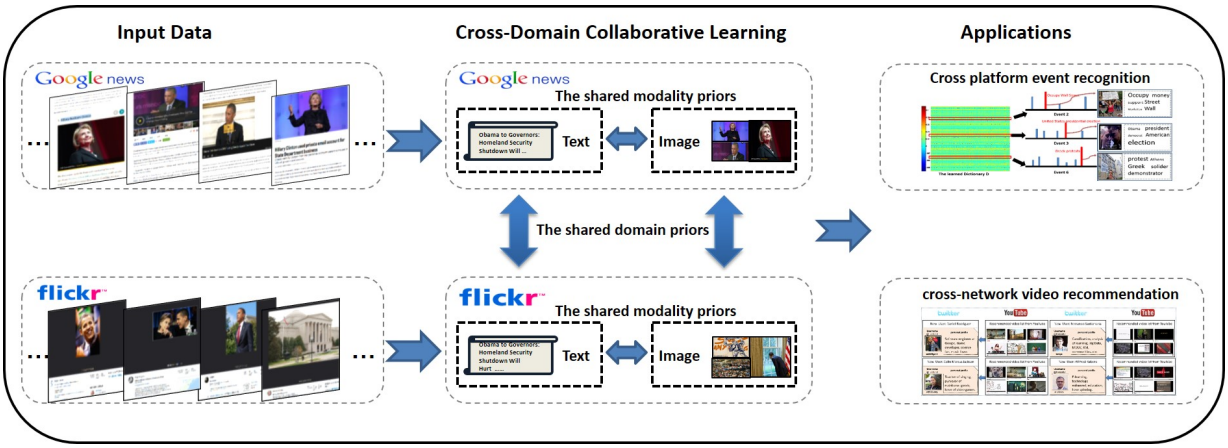


Figure 2: Illustration of the key idea of our cross-domain collaborative learning. For simplicity, we only show an example for cross-platform data association. Here, there are two domains (Google News and Flickr) with two modalities (Text and Image)(All photos via Flickr under Creative Commons License).

s Twitter and YouTube, there are many social links by users. For each user, he/she will have video-related behaviors on YouTube and tweeting behaviors on Twitter, and we explore the overlapped user account linkage between Twitter and YouTube to realize cross-network user association. Even though the shared domain information can help bridge the domain gap, it is still a challenging problem to find the most effective way to explore the useful information, as the data are inherently heterogeneous, noisy and ambiguous.

Recently, how to explore useful information via cross domain collaborative learning in social multimedia has attracted much research interest. Basically, there are two major research topics: (1) cross-domain feature learning in multimedia by adopting a cross-domain constraint to make different domains share a common feature space [1, 4, 5]. For example, Yang et al. [1] propose a cross-domain feature learning algorithm based on stacked denoising auto-encoders, and apply the learned features for three important applications: sentiment classification, spam filtering, and event classification. (2) personalization recommendation in social links by modeling cross-network user behaviors [6, 7]. For example, a cold-start recommendation solution is proposed by aggregating user profiles in Flickr, Twitter, and Delicious [6]. Yan et al. investigate into cross-network social relation and behavior information to address the cold-start friend recommendation problem [7]. These two topics focus on feature learning and user modeling in cross domain analysis by combining the virtues of different domains to complement each other. They are both challenges and benefits. (1) The media data (e.g., event topics, social links) have **multi-domain** property. As shown in Figure 1, the media data can come from multiple domains (e.g., Flickr, Google News, YouTube, and Twitter), and they can complement each other, but also have domain discrepancy. (2) The media data have **multi-modal** property. As shown in Figure 1(a), each data instance in social media sites can be described with images and texts simultaneously, and the textual and visual information can also complement each other. (3) The media data have **sparse** property. For example, the social behaviors of the users in one social network are sparse. In particular, most users have no chance to browse or review most images and videos. When a new user registers on Youtube, the system knows nothing about his/her interactions on videos and cannot conduct video recommendation.

In this paper, motivated by the previous work, we propose a novel generic Cross-Domain Collaborative Learning (CDCL) framework based on non-parametric Bayesian dictionary learning model for cross-domain data analysis. In our CDCL, the non-parametric Bayesian dictionary learning model can explore the multi-domain, multi-modality, and sparse properties jointly. (1) To deal with the domain discrepancy, we adopt the shared domain priors across multiple domains to make them share a common feature space. (2) To make use of the multi-modal property, we learn the sparse representation of multi-modal data by introducing the shared modality priors to infer the sparse structure shared among different modalities of media data. (3) To deal with the sparsity of the media data, we learn the shared dictionary space to bridge cross-domain information. Due to the sparsity of user behavior in one social network, we exploit social relations and behaviors of users in the auxiliary social network to help estimate their preferences on other social network by learning the shared dictionary space, which can perform the cold-start recommendation task. The details of our CDCL algorithm is shown in Figure 2. For simplicity, we only show an example for cross-platform data association. There are two domains (Google News and Flickr) with two modalities (Text and Image) related to the event "United States Presidential Election". In the left panel of Figure 2, the related data associated with the event include textual and visual information. Here, each social event instance contains text and its corresponding images. Since the multi-modal data among different domains have their own characteristics but also have their commonalities, we can collaboratively learn the shared feature representation by adopting the shared domain priors and modality priors across multiple domains as shown in Figure 2. As a result, the proposed CDCL can effectively combine the virtues of different information sources to complement each other for cross-domain multi-modal data analysis. The proposed generic framework can be applied for many applications, such as cross-platform event recognition and cross-network video recommendation. The cross-platform event recognition is to use multi-modal data from multiple domains to conduct social event recognition. And the cross-network video recommendation is to leverage users' rich cross-network activity data to help estimate their preferences on other social platforms. For example, by using the overlapped user account linkage between Twitter and YouTube, and considering both the Twitter tweeting activities and historical interactions with YouTube videos, we design a cold-start recommendation task for the new YouTube user by the proposed CDCL method. We evaluate our method on these two applications and the results demonstrate its effectiveness in social multimedia. Compared with the existing methods, the contributions of this work are threefold.

- We propose a generic cross-domain collaborative learning framework based on non-parametric Bayesian dictionary learning model for cross-domain data analysis, and the proposed CDCL can effectively make use of the virtues of different information sources to complement and enhance each other.
- The proposed non-parametric Bayesian dictionary learning model can effectively adopt the shared domain and modality priors to collaboratively learn the shared feature representation to deal with the domain discrepancy with considering the multi-modal property.
- We evaluate the proposed CDCL method on two different applications in social multimedia and demonstrate that it achieves much better performance than existing methods. Besides, we collect a large-scale dataset for research on multi-modality cross-domain social event analysis, and will release it for academic use.

The rest of the paper is organized as follows. In Section 2, the related work is reviewed. Section 3 introduces the formulation of the CDCL. Two cross-domain applications are presented in Section 4. In Section 5, we report and analyze extensive experimental results. Finally, we conclude the paper with future work in Section 6.

2. RELATED WORK

In this Section, we briefly review previous methods which are most related to our work including cross-domain feature learning and cross-network collaborative learning in multimedia.

Cross-domain Feature Learning: In cross-domain feature learning, most of the existing methods aim to propagate the knowledge from an auxiliary domain to a target domain, which can learn feature representation by using data from one domain space to enhance the learning tasks of other domain spaces. In these methods, the auxiliary domain can be considered as the prior knowledge and experience guidance to perform new learning task on target domain. To achieve this goal, many methods adopt a cross-domain constraint to make different domains share a common feature space [1, 4, 5, 8]. Blitzer et al. [4] introduce structural correspondence learning to automatically induce correspondences among features from different domains by modeling their relations with pivot features that appear frequently in both domains. In [5], it reduces the distance across two domains by learning a latent feature space where domain similarity is measured through maximum mean discrepancy. With the explosive growth of multi-media data on the Web, cross-media learning [9, 10, 11, 12, 13] also has drown much attention in the past few years. Bian et al. [12] propose the CM-LDA method to model the relations among different media types by introducing a shared latent variable Z. Yang et al. [9] integrate semi-supervised learning and transfer learning techniques to exploit manually-labeled images for video tagging.

Different from the existing methods, the proposed CDCL algorithm adopts a shared dictionary space learning strategy to bridge different domains. There are some existing dictionary learning approaches. In [14], a novel coupled dictionary training method is proposed for single image super-resolution based on patch-wise sparse recovery, where the learned couple dictionaries can connect the low with high-resolution image patch spaces. Semi-coupled dictionary learning model is proposed to solve such cross-style image synthesis problems [15]. In these dictionary learning methods, the sparse coefficient is estimated by assuming the reconstructed residual error or the sparsity level. However, we usually do not know the residual error or sparsity level. If the settings do not agree with the ground truth, the performance can significantly degrade. Instead, Zhou et al. [16] and Yuan et al. [17] introduce a non-parametric Bayesian model to address these problems. In [16], the dictionary learning method with the non-parametric beta process is presented, where the beta process is employed as a prior for learning the dictionary, and this non-parametric method can naturally infer an appropriate dictionary size. In [17], a novel multi-task sparse learning model is proposed for human action recognition. In these two methods, the non-parametric Bayesian methods perform well in image denoising, image inpainting, compressive sensing, and human action recognition. In this paper, motivated by the previous work, we propose a novel non-parametric Bayesian dictionary learning model for cross-domain data analysis by using the shared domain and modality priors in social multimedia. Different from the proposed method, in [16], it mainly adopts beta process prior to model the image's sparse structure information and does not consider the cross-domain information. While our goal is to model cross-domain information with the shared domain and modality priors to complement and enhance each other.

Cross-network Collaborative Learning: The cross-network collaborative learning has recently attracted broad attentions. Most of the existing methods are devoted to taking advantage of different social networks' information towards collaborative applications. Suman et al. [18] exploit the real-time and socialized characteristics of the Twitter tweets to facilitate video applications on YouTube. Abel et al. [19] investigate tag profiles for the same user on Flickr, Twitter and Delicious, and discover consistency and replication characteristics in cross-platform user behavior.

The challenge in cross-network collaborative learning is how to bridge the domain gap by the shared content information, which is to make the strengths of one domain complement the weaknesses of the others. In terms of cross-network collaborative learning in social media, our work is related to [7] and [20]. In [7], the cross-network social relation and behavior information are adopted to address the cold-start friend recommendation problem. Different from [7], our work focuses on applying cross-network collaborative learning method to a YouTube video recommendation applications to deal with the cold-start problem. This work is also different from [20] which uses coupled dictionary learning method to conduct the cross-network topic association to meet the YouTube video promotion demand. Instead of the coupled dictionary learning method, we introduce the non-parametric Bayesian dictionary learning model by using the shared domain priors to collaboratively learn the shared dictionary space. Moreover, the proposed CDCL method is a generic framework for cross-domain data analysis, and the cross-network collaborative learning is only one of our applications in social multimedia.

3. OUR APPROACH

In this Section, we will first introduce the details of the proposed cross-domain collaborative learning algorithm, and then show its model inference.

3.1 Cross-Domain Collaborative Learning

The cross-domain collaborative learning is to explore the virtues of different information sources to complement and enhance each other for cross-domain data analysis. To achieve this goal, we propose a generic collaborative learning framework via non-parametric Bayesian dictionary learning model. In the proposed model, it can effectively explore the multi-domain property and the multi-modality property of cross-domain data to collaboratively learn the shared feature representation to deal with the domain discrepancy and help bridge the domain gap. For cross-domain data analysis, without loss of generality, we can assume that the data have J do-

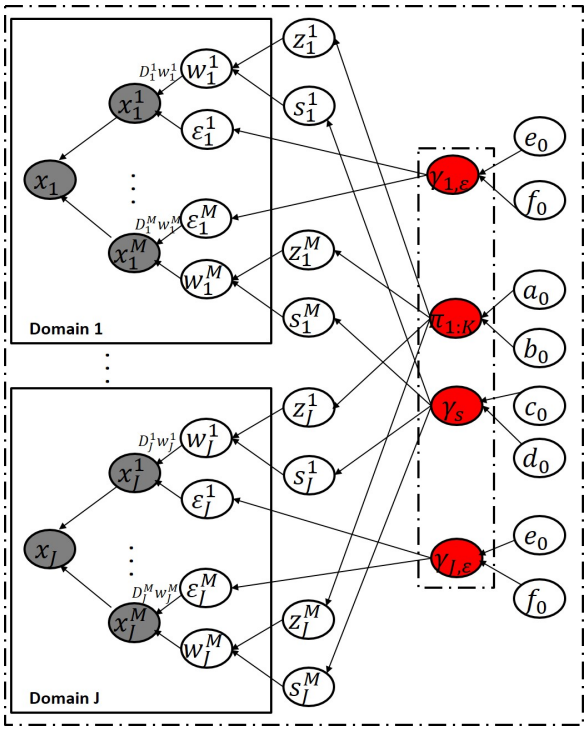


Figure 3: The graphical representation of our cross-domain collaborative learning algorithm. The red circles represent the shared priors to associate with the relevant information and collaboratively learn the shared feature space in different domains. For details, please refer to the corresponding text in Section 3.1.

mains with M modalities. Let $\mathbf{X} = [\mathbf{x_1}, \ldots, \mathbf{x_j}, \ldots, \mathbf{x_J}]$ denote the data instances in J domains. Here, $\mathbf{x_j} = [\mathbf{x_j^1}, \cdots, \mathbf{x_j^M}]$ denote M modalities in the j-th domain, and $\mathbf{x_j^m} \in \mathbf{R}^{\mathbf{n_j^m}}$. The n_j^m is the dimensionality of the feature of the m-th modality in the j-th domain. In this paper, these instances $\mathbf{X}$ can be either social event data or user information, such as social events described by images and texts or the user account linkages between Twitter and YouTube.

To model these data, sparse representation has shown encouraging performance in traditional methods [21]. Given an instance $\mathbf{x_j^m}$ in the j-th domain with the m-th modality, it can be represented as a linear combination of elements in dictionary $\mathbf{D}_j^m$ with an error term ε_j^m as shown in Eq.(1).

$$\mathbf{x}_j^m = \mathbf{D}_j^m \mathbf{w}_j^m + \varepsilon_j^m \tag{1}$$

Here, the columns of the matrix $\mathbf{D}_j^m \in \mathbf{R}^{n_j^m \times K}$ represent the K elements of the dictionary, the $\mathbf{w}_j^m$ is the sparse feature coefficient, and the ε_j^m represents measurement noise with the m-th modality in the j-th domain. The sparse representation in Eq.(1) only models an instance with single modality in one domain. To improve it for modeling instance with multi-modality in cross-domain, we should take into account the following issues: (1) How to make use of the multi-modality and multi-domain properties to learn cross-domain data representations; (2) How to make the coefficient $\mathbf{w}_j^m$ have a sparse constraint; (3) How to learn a shared dictionary space $\mathbf{D}_j^m$ to bridge the domain gap; To deal with the above issues, we propose an effective cross-domain collaborative learning algorithm by introducing the non-parametric Bayesian dictionary learning model as shown in Figure 3. In the proposed model, (1) To deal with the domain discrepancy, we add the the shared domain priors π, γ_s to associate with the information across multiple domains. Meanwhile, the shared modality priors $\gamma_{j,\varepsilon}, \pi_j, \gamma_{j,s}$ are adopted to associate with the multi-modality information in the j-th domain. (2) The sparseness is achieved by introducing the Beta Process priors. Since different domains may favor different sparse reconstruction coefficients, the constraint of joint sparsity across different domains is necessary, which is to enforce the robustness in the sparse coefficient estimation. Due to the shared priors, our model can realize joint sparsity across different domains. (3) The shared dictionary space is learned by using the shared domain and modality priors, and it can bridge the domain gap for cross-domain data analysis.

In the non-parametric Bayesian dictionary learning model, the beta process is employed as a prior to learn the dictionary $\mathbf{D}_j^m$, and the number of dictionary elements across all domains and their relative importance can be inferred non-parametrically. As a result, the sparseness of the coefficient can be achieved via the Beta process priors rather than the computationally expensive ℓ_1 norm penalty. The Beta process (BP) is developed in [22], and the BP with parameters $a_0 > 0$, $b_0 > 0$, and base measure H_0, is represented as $BP(a_0, b_0, H_0)$. The stick-breaking construction of a beta process $H \sim BP(a_0, b_0, H_0)$ is represented as:

$$H(\psi) = \sum_{k=1}^{K} \pi_k \delta_{\psi_k}(\psi) \tag{2}$$

Here, $\pi_k \sim Beta(a_0/K, b_0(K-1)/K)$ and $\psi_k \sim H_0$. The $H(\psi)$ represents a vector of K probabilities, with each associated with a respective atom ψ_k, and ψ_k is the atom distributed according to H_0. When $K \to \infty$, $H(\psi)$ corresponds to an infinite-dimensional vector of probabilities, and each probability has an associated atom ψ_k drawn i.i.d. from H_0. In the proposed model, we set the sparse feature coefficient $\mathbf{w}_j^m = \mathbf{z}_j^m \odot \mathbf{s}_j^m$, where $\odot$ represents the Hadamard (element-wise) multiplication of two vectors, the binary vector $\mathbf{z}_j^m \in \{0,1\}^K$ denotes which of the K columns of D_j^m are used to represent the instance $\mathbf{x}_j^m$, and the weights $\mathbf{s}_j^m \sim N(0, \gamma_s^{-1}\mathbf{I}_K)$ are introduced to impose that the reconstruction coefficients of the dictionary are not always binary, where γ_s is the precision or inverse variance. That is to say, let the atoms ψ_k correspond to the candidate members of our dictionary D_j^m, and the k-th component of the binary vector $\mathbf{z}_j^m$ is drawn $\mathbf{z}_{j,k}^m \sim \text{Bernoulli}(\pi_k)$. As shown in Figure 3, the shared priors $\pi, \gamma_s, \gamma_\varepsilon$ are utilized to learn the shared dictionary space and the feature representation of instances across multiple domains. The hierarchical form of the cross-domain collaborative learning method can be described as:

$$\begin{aligned} &\mathbf{x}_j^m = \mathbf{D}_j^m \mathbf{w}_j^m + \varepsilon_j^m, m = 1, \ldots, M, j = 1, \ldots, J \\ &\mathbf{D}_j^m = [d_{j,1}^m, \cdots, d_{j,K}^m] \\ &\mathbf{w}_j^m = \mathbf{z}_j^m \odot \mathbf{s}_j^m \\ &\varepsilon_j^m \sim N(0, \gamma_{j,\varepsilon}^{-1}\mathbf{I}_{n_j^m}), \end{aligned} \tag{3}$$

where $d_{j,k}^m \sim N(0, n_j^{m(-1)}\mathbf{I}_{n_j^m})$, $\mathbf{z}_j^m \sim \prod_{k=1}^{K} \text{Bernoulli}(\pi_k)$, $\pi_k \sim \text{Beta}(a_0/K, b_0(K-1)/K)$, $\mathbf{s}_j^m \sim N(0, \gamma_s^{-1}\mathbf{I}_K)$, $\gamma_s \sim \Gamma(c_0, d_0)$, and $\gamma_{j,\varepsilon} \sim \Gamma(e_0, f_0)$. The gamma hyper-priors placed on γ_s and $\gamma_{j,\varepsilon}$ are typically non-informative. Note that, for simplicity, independent conjugacy Gaussian priors for $d_{j,k}^m$, $\mathbf{s}_j^m$, and ε_j^m are adopted. As a result, the proposed CDCL method can effectively not only make use of the information of each domain, but also combine the virtues of other domains to complement and enhance each other by the shared domain and modality priors.

3.2 Model Inference

In the proposed CDCL method, the full likelihood probability can be factorized as:

$$P(X, D, Z, S, \pi, \gamma_s, \gamma_\varepsilon) = \prod_{j=1}^{J}\prod_{m=1}^{M} N(x_j^m; D_j^m(s_j^m \odot z_j^m), \gamma_{j,\varepsilon}^{-1} I_{n_j^m}) \prod_{j=1}^{J}\prod_{m=1}^{M}\prod_{k=1}^{K} N(d_{j,k}^m; 0, n_j^{m(-1)} I_{n_j^m}) \text{Bernoulli}(\mathrm{z}_{j,k}^m; \pi_k) \prod_{k=1}^{K} \text{Beta}(\pi_k; a_0, \mathrm{b}_0) \prod_{j=1}^{J} \Gamma(\gamma_{j,\varepsilon}; e_0, \mathrm{f}_0)\Gamma(\gamma_s; c_0, \mathrm{d}_0). \quad (4)$$

An exact model inference is often intractable in many non-parameter Bayesian models, and some appropriate methods must be used, such as variational inference [23] and Gibbs sampling [24]. To estimate the latent variables conditioned on the observed variables, namely $\mathbf{D}$, $\mathbf{Z}$, $\mathbf{S}$, π, γ_s, γ_ε, we employ Gibbs sampling method to obtain samples of latent variables and estimate unknown parameters in our model. In a Gibbs sampler, it iteratively samples new assignments of latent variables by drawing from the distributions conditioned on the previous state of the model. We list the update rules for latent variables $\mathbf{D}$, $\mathbf{Z}$, $\mathbf{S}$, π, γ_s, γ_ε as follows:

We first sample the dictionary variable $\mathbf{D}_j^m = [d_{j,1}^m, \cdots, d_{j,K}^m]$ according to the posterior probability as shown in Eq.(5).

$$P(d_{j,k}^m|-) \propto N(x_j^m; D_j^m(s_j^m \odot z_j^m), \gamma_{j,\varepsilon}^{-1} I_{n_j^m}) N(d_{j,k}^m; 0, n_j^{m(-1)} I_{n_j^m}) \quad (5)$$

Here, the $d_{j,k}^m$ can be drown from a normal distribution $p(d_{j,k}^m|-) \sim N(u_{d_{j.k}^m}, \Sigma_{d_{j.k}^m})$.

Then, we sample the binary vector $\mathbf{z}_j^m = [z_{j,1}^m, \cdots, z_{j,K}^m]$ according to the following posterior probability,

$$P(z_{j,k}^m|-) \propto N(x_j^m; D_j^m(s_j^m \odot z_j^m), \gamma_{j,\varepsilon}^{-1} I_{n_j^m}) \text{Bernoulli}(\mathrm{z}_{j,k}^m; \pi_k) \quad (6)$$

Here, when $z_{j,k}^m = 1$, the $P_1 \propto N(x_j^m; D_j^m(s_j^m \odot z_j^m), \gamma_{j,\varepsilon}^{-1} I_{n_j^m}) \cdot \pi_k$; when $z_{j,k}^m = 0$, the $P_0 = 1 - \pi_k$. We can draw $z_{j,k}^m$ according to the Bernoulli distribution $z_{j,k}^m \sim \text{Bernoulli}(\frac{\mathrm{P}_1}{\mathrm{P}_1 + \mathrm{P}_2})$.

Next, we sample the weight variable $\mathbf{s}_{j,k}^m = [s_{j,1}^m, \cdots, s_{j,K}^m]$ as in Eq.(7).

$$P(s_{j,k}^m|-) \propto N(x_j^m; D_j^m(s_j^m \odot z_j^m), \gamma_{j,\varepsilon}^{-1} I_{n_j^m}) N(s_j^m; 0, \gamma_s^{-1} I_K) \quad (7)$$

Here, as $d_{j,k}^m$, the $s_{j,k}^m$ can be drown from a normal distribution $p(s_{j,k}^m|-) \sim N(u_{s_{j,k}^m}, \Sigma_{s_{j.k}^m})$.

Finally, we sample the shared priors $\pi, \gamma_s, \gamma_\varepsilon$ with the following updating rules.

$$P(\pi_k|-) \propto \text{Beta}(\pi_k; a_0, b_0) \prod_{j=1}^{J}\prod_{m=1}^{M} \text{Bernoulli}(z_{j,k}^m; \pi_k) \quad (8)$$

$$P(\gamma_s|-) \propto \Gamma(\gamma_s; c_0, d_0) \prod_{j=1}^{J}\prod_{m=1}^{M} N(s_j^m; 0, \gamma_s^{-1} I_K) \quad (9)$$

$$P(\gamma_{j,\varepsilon}|-) \propto \Gamma(\gamma_{j,\varepsilon}; e_0, f_0) \prod_{j=1}^{J}\prod_{m=1}^{M} N(x_j^m; D_j^m(s_j^m \odot z_j^m), \gamma_{j,\varepsilon}^{-1} I_{n_j^m}) \quad (10)$$

Algorithm 1 The proposed CDCL method for cross-platform event recognition.

Input: data in auxiliary domain D_a; training and testing data in target domain D_t; Iteration number T_{gibbs};
Output: predict class labels for testing documents in D_t.

// Learn the shared domain priors in auxiliary domain D_a.
Initialize the dictionary variable via the K-SVD
Initialize latent variables $\mathbf{z}, \mathbf{s}, \pi, \gamma_s, \gamma_\varepsilon$
for $t := 1 \rightarrow T_{gibbs}$ **do**
  Run Gibbs sampling strategy for all instances in the auxiliary domain D_a according to Eq.(5) $\sim$ Eq.(10)
end for
// Predict class labels for testing samples in D_t
Initialize the domain priors with the learned values in D_a
Learn the sparse representation $\mathbf{w}$ for all instances in D_t
Predict class labels of testing data using Linear SVM.

Note that, the priors π, γ_s are shared across multiple domains, while the prior γ_ε is only shared across multiple modalities in a single domain.

4. APPLICATIONS

In this Section, we introduce how to leverage our generic model for two cross-domain applications: *cross-platform event recognition* and *cross-network video recommendation*.

4.1 Cross-platform Event Recognition

The cross-platform event recognition is to make use of the virtues of different domains to learn the shared feature representation for social event recognition. A popular social event that is happening around us can spread very fast. As a result, there are a large amount of social events with multi-modality (e.g., images, videos, and text) in many different domains (e.g., Flickr and Google News). Therefore, it is important to automatically identify and recognize the interesting social events from massive social media data. The critical challenge is how to make use of the cross-domain multi-modality social event data. The social event recognition is normally studied with textual features. In addition to textual information, social events also have rich visual information. For an event in different sites, it may have different textual descriptions (comments, tags, etc.) due to different users. However, it may have very similar visual information, such as images or videos, which are useful for social event recognition across time and sites. For example, the event "United States Presidential Election", its stories with the picture of Obama are highly related. Therefore, multi-modality fusion is useful for social event recognition. In addition, different platforms can complement and enhance each other. For example, most events on Google News are from the official, but they have a lot of comments by web users on Flickr. The images on Google News are generally captured by professional journalists while the images on Flickr are typically captured by common users, who may upload images which do not focus on the targets of a specific event perfectly. Therefore, social events in different domains can help each other, especially when the strengths of one domain complement the weaknesses of the other.

Next, we will introduce how to apply the proposed CDCL algorithm for the cross-platform event recognition by considering two domains (Google News and Flickr) with two modalities (text and image). Here, each social event has many documents, and each document is an event instance (event sample) including text and images. Our goal is to classify each document by using the informa-

tion from Google News and Flickr. In general, these two domains can be viewed as auxiliary domain and target domain, respectively. The auxiliary domain can be modeled as the prior knowledge and experience to perform new learning task on target domain, which can be facilitated to improve the classification accuracies, especially when the size of available labeled training samples on target domain is not large enough. The overall procedure for cross-domain social event recognition is depicted in Algorithm 1. We first learn the shared domain priors of the proposed CDCL model by alternately sampling instances from the auxiliary domain. Here, the data from the auxiliary domain are used to infer the shared domain and modality priors $\pi, \gamma_s, \gamma_\varepsilon$. Then, we use the learned priors to train the proposed CDCL model and learn the sparse feature representation $\mathbf{w}$ for all instances in the target domain D_t. Then, the Linear SVM [25] is trained and used to predict class labels of the testing data in D_t. Here, a half of instances in the target domain are randomly sampled to train the SVM classifiers, and the other half of the instances are used for testing.

4.2 Cross-network Video Recommendation

The cross-network video recommendation is to leverage users' rich cross-network activity data to help estimate their preferences on other social network, especially for a new user or the users with few records. In this paper, we exploit social relations and behaviors of users in the auxiliary social platform, and apply our CDCL method to help another platform conduct the cold-start recommendation task. Specifically, we employ Twitter and YouTube as the two platforms, which are connected due to the overlapped user account linkage between Twitter and YouTube. When a new user registers on Youtube, the system knows nothing about his/her interactions on the videos and cannot conduct video recommendation. By considering both the Twitter tweeting activities and historical interactions with YouTube videos, the proposed CDCL algorithm can present a YouTube video recommendation application to deal with the cold-start problem.

In cross-network video recommendation, each user $u \in U$ can be represented as a 2-dimensional tuple $< \mathbf{u}^T, \mathbf{u}^Y >$ when given a set of overlapped users U, where the user u is considered as an instance, the $\mathbf{u}^T$ represents the user' tweet information in Twitter, and the $\mathbf{u}^Y$ represents that the user u has interacted with the videos in Youtube. Here, we consider each user's tweet history information as one document and employ the standard Latent Dirichlet Allocation to the corpus composed by all the Twitter users. Feature description of $\mathbf{u}^T$ can be represented as $\mathbf{u}^T = \{u_1^T, \cdots, u_{K_T}^T\}$, where K_T is the number of topics in the latent Twitter topic space. Similarly, YouTube user's feature representation $\mathbf{u}^Y$ can be obtained with his/her interested video set. We consider all the video set as the feature items of the YouTube users and take the commonly used vector space model to represent the YouTube user. The $\mathbf{u}^Y$ can be represented as $\mathbf{u}^Y = \{u_1^Y, \cdots, u_{K_Y}^Y\}$, where K_Y is the number of YouTube videos in our dataset, and $u_n^Y = 1$ denotes that the YouTube user u has been interested in the n-th YouTube video and $u_n^Y = 0$ otherwise. The task of cross-network video recommendation is, for a given new user $u \in U$ on YouTube, to recommend a ranking list of videos V_u according to the user' interest by considering the user' tweet activities $\mathbf{u}^T$ on Twitter.

The proposed cross-network video recommendation solution consists of two steps, i.e., cross-network dictionary learning and video recommendation for the new YouTube users, as shown in Algorithm 2. At the first step, with the obtained Twitter and YouTube user feature representation $\mathbf{u}^T, \mathbf{u}^Y$, we utilize the proposed CDCL method to discover the shared latent structure among different networks. Specifically, given the user feature representation

Algorithm 2 The proposed CDCL method for cross-network video recommendation.

Input: User representation $\mathbf{u}^T \in U$ on Twitter; User representation $\mathbf{u}^Y \in U$ on YouTube; Candidate YouTube videos $v_t \in V$; A test user $u_t \in U_t$ on YouTube.
Output: A ranked list of videos V_u for u_t.

// Do the cross-network dictionary learning
1: Learn the shared dictionary space $\mathbf{D}^T = \{d_1^T, \cdots, d_K^T\}$ on Twitter according to Eq.(5) $\sim$ Eq.(10)
2: Learn the shared dictionary space $\mathbf{D}^Y = \{d_1^Y, \cdots, d_K^Y\}$ on YouTube according to Eq.(5) $\sim$ Eq.(10)
//Recommend videos for a new YouTube user u_t
3: Obtain $\mathbf{u}^T \in R^{K_T \times 1}$ by the user's tweet history information
4: Estimate the corresponding sparse coefficient $\mathbf{w}$ by Eq.(11)
5: Obtain the feature representation $\mathbf{u}^Y$ via Eq.(12)
6: Recommend a ranked list of videos V_u for u_t by Eq.(13)

$\mathbf{u}^T = \{u_1^T, \cdots, u_{K_T}^T\}$ and $\mathbf{u}^Y = \{u_1^Y, \cdots, u_{K_Y}^Y\}$, we can learn the shared dictionary space for the users across different domains and obtain the dictionary variables $\mathbf{D}^T = \{d_1^T, \cdots, d_K^T\}$ and $\mathbf{D}^Y = \{d_1^Y, \cdots, d_K^Y\}$ on Twitter and YouTube, respectively. With the derived shared dictionary spaces $\mathbf{D}^T$ and $\mathbf{D}^Y$, we are able to realize the transfer of user's feature distribution between different networks. Therefore, we can conduct video recommendation for new user on YouTube by his/her tweet history information in Twitter. The second step is to recommend a ranking list of videos V_u for a given new user in YouTube based on the learned $\mathbf{D}^T$ and $\mathbf{D}^Y$. Specifically, given a new user on YouTube, we can obtain the user's feature representation $\mathbf{u}^T \in R^{K_T \times 1}$ on Twitter by his/her tweet history information. Then, we can estimate the user's sparse feature coefficient $\mathbf{w}$ via Eq.(11) for $\mathbf{u}^T$.

$$\mathbf{u}^T = \mathbf{D}^T \mathbf{w} + \varepsilon_j \tag{11}$$

Since the unique user across different domains has the shared dictionary space ($\mathbf{D}^T$ and $\mathbf{D}^Y$) and sparse feature coefficient ($\mathbf{w}$), we can adopt these learned parameters to help conduct the transfer of user's feature distribution between the Twitter and YouTube. Therefore, the new user's YouTube feature representation can be defined as:

$$\mathbf{u}^Y = \mathbf{D}^Y \mathbf{w} \tag{12}$$

As a result, given the new user $\mathbf{u}^Y$ and candidate YouTube videos $\mathbf{v}_t \in V$ represented in the same feature space, the recommended videos from YouTube are ranked by Eq.(13).

$$sim(\mathbf{u}^Y, \mathbf{v}_t) =< \mathbf{u}^Y, \mathbf{v}_t >= \sum_{k=1}^{K_T} u_k^T \cdot v_{k,t} \tag{13}$$

5. EXPERIMENTS

In this Section, we evaluate the performance of the proposed CDCL algorithm on two different applications: cross-platform event recognition and cross-network video recommendation. The extensive results demonstrate the effectiveness of our CDCL algorithm for cross-domain collaborative learning in social multimedia.

5.1 Cross-platform Event Recognition

5.1.1 Dataset Collection

For social event recognition, the evaluation dataset is constructed from online social platforms. Nowadays, there are already some

Table 1: Illustration of the event name, duration time, and number of documents for each event in our collected social event dataset.

Event ID	Event Name	Start Time	End Time	Google News		Flickr	
				#Images	#Text	#Images	#Text
1	Senkaku Islands dispute	2008.06	2012.12	3743	2495	6617	6617
2	Occupy Wall Street	2011.09	2012.09	5601	3108	7151	7151
3	United States Presidential Election	2009.10	2013.01	5169	3446	7352	7352
4	War in Afghanistan	2001.10	2012.08	5373	2915	7172	7172
5	North Korea nuclear program	2000.01	2012.04	3969	2640	8635	8635
6	Greek protests	2011.05	2012.04	3900	2630	7385	7385
7	Mars Reconnaissance Orbiter	2005.04	2012.08	3901	2600	7188	7188
8	Syrian civil war	2011.01	2013.01	4899	3266	7426	7426

Table 2: The event classification accuracy of different methods.

Methods	Accuracy	
	#Google News	#Fickr
BOW	0.797	0.857
CCA	0.758	0.861
SRC-L1	0.820	0.857
SRC-L1-DL	0.843	0.862
CDCL-s	0.834	0.861
CDCL-c	**0.848**	**0.877**
CDCL	**0.876**	**0.885**

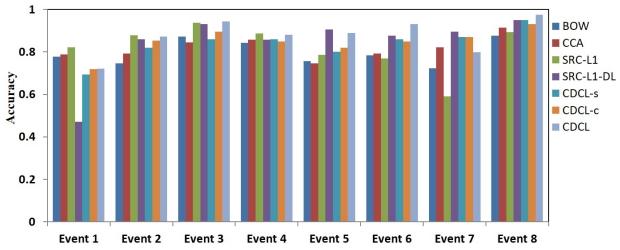


Figure 4: The classification accuracy for each event on Google News.

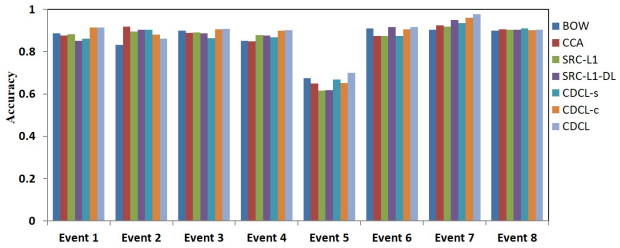


Figure 5: The classification accuracy for each event on Flickr.

public event datasets, such as the MediaEval social event detection(SED) [26]. However, the existing MediaEval SED dataset include only social media content created by people and do not have current hot social events. Besides, the existing MediaEval SED dataset do not have multi-modality cross-domain information. To analyze event data with the multi-modal and multi-domain properties, we mainly focus on 8 complex and public social events happened in the past few years, and collect the dataset by ourselves from Google News and Flcikr. For these 8 events, we manually create the introduction page of each event or download it from the Wikipedia page[1], which contains the whole stories of each event. We then search and download related text and its corresponding images from Google News and Flickr based on the keywords in the whole timeline of each social event. The detail of our collected dataset is shown in Table 2. The collected 8 social events cover a wide range of topics including politics, economics, military, society, and so on. For each social event, there are about 2000 to 9000 documents including text and its corresponding images.

5.1.2 *Feature Extraction*

For textual description, we use stemming method and stop words elimination and remove words with a corpus frequency less than 15 in the whole stories of the event, and take the commonly used vector space model to represent the textual information. For visual description, we adopt the popular sparse coding method [27]. In our implementation, we densely sample SIFT points from images, and adopt K-means to build a codebook. For each SIFT point, the Localized Soft-assigment Coding (LSC) is adopted to obtain its descriptor. Then, the max pooling and the Spatial Pyramid Matching (SPM) strategy are adopted to obtain image representation.

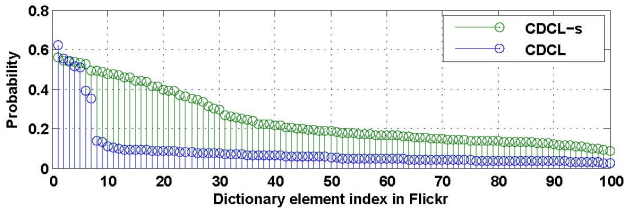
5.1.3 *Results and Analysis*

In our experiment setting, the K is set to 100. Note that not all K dictionary elements are used in the model. The number of shared dictionary elements will be determined by the shared domain priors. The hyperparameters within the gamma distributions are set as $c_0 = d_0 = e_0 = f_0 = 10^{-6}$ as in [16]. Since the data overwhelms these prior values when calculating posterior distributions, our algorithm is robust to the initial values. We set $T_{gibbs} = 100$ and use the results of the last iteration.

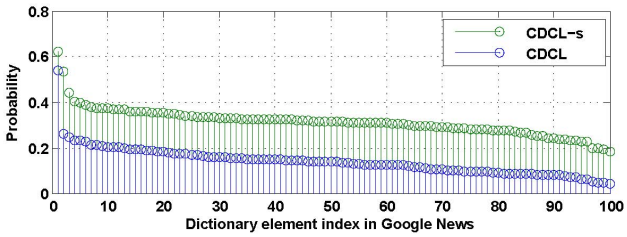
To demonstrate the effectiveness of the proposed CDCL model for cross-domain event analysis, we compare it with the most related baseline methods (BOW, CCA, SRC-L1, SRC-L1-DL):

- BOW: It is to concatenate the textual and visual features as discussed in Section 5.1.2 to represent each event document.
- Canonical Correlation Analysis (CCA) [28]: The CCA is a classical method in cross modal retrieval by learning a common space across multi-modal data. The text and image features are obtained by the maximally correlated subspace.
- SRC-L1: It is to adopt the sparse representation based on traditional ℓ_1 regularization and the dictionary is learned via the traditional K-SVD method.
- SRC-L1-DL: It is to adopt the sparse representation based on traditional ℓ_1 regularization and the dictionary is learned by the proposed non-parametric Bayesian model with the auxiliary domain as in Algorithm 1.

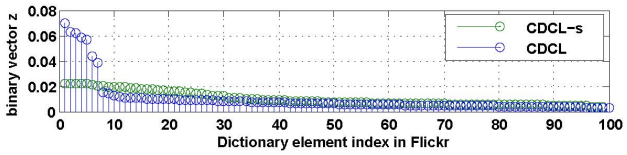
[1] http://www.wikipedia.org



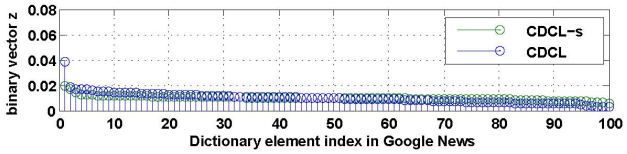
(a) Dictionary weights π_k (Flickr)



(b) Dictionary weights π_k (Google News)



(c) Statistics of binary vector z (Flickr)

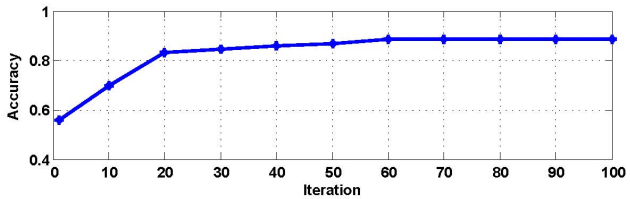


(d) Statistics of binary vector z(Google News)

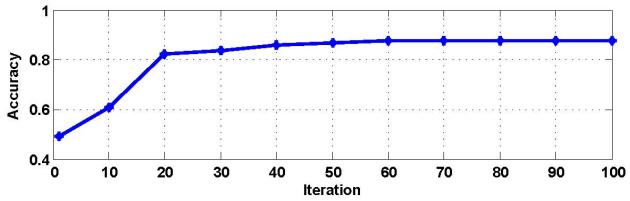
Figure 6: The shared dictionary learning results by CDCL-s and CDCL. (a-b) the comparison of dictionary weights π_k on Flickr and Google News, respectively. (c-d) the statistics results of binary vector z on Flickr and Google News, respectively.

For the cross-platform event analysis, our goal is to classify the event instances on Google News/Flickr domain with the help of the auxiliary domain Flickr/Google News, respectively. With different experimental settings, we have 3 methods CDCL-s, CDCL-c, and CDCL. The CDCL-s is only trained in the single domain without the help of the auxiliary domain, and the shared priors is initialized with random values. The CDCL-c is trained with the help of the auxiliary domain, but ignores the multi-modal constraint by directly concatenating the textual and visual features. The CDCL is trained by using the auxiliary domain with multi-modal property to obtain the shared domain priors and modality priors. After learning the representations, the Linear SVM is utilized as the classifier.

The classification results of different methods are shown in Table 2, and the accuracy comparison of each event class is given in Figure 4 and Figure 5. Based on these results, we have the following observations. (1) The BOW model shows inferior classification performance. This is because the BOW models textual and visual words obscurely and cannot differentiate the associations between multi-modal data. (2) The CCA and our CDCL achieve better performance than the BOW, which shows that it is useful to model and fuse the textual and visual information. (3) The SRC-L1-DL achieves better average classification accuracy than the SRC-L1, which shows that the dictionary learning method by adopting the auxiliary domain can obtain a more compact and representative dictionary to improve the average performance. (4) Overall, the proposed CDCL method consistently outperforms other existing methods based on the average classification accuracy. The major reason is that the proposed non-parametric Bayesian dictionary learning model can adopt the shared domain priors and modality priors to collaboratively learn the feature representation by considering the domain discrepancy and the multi-modal property. We also observe that our CDCL method is much worse than the SRC-L1 for the event 1 in Figure 4 and the event 2 in Figure 5, respectively. This is may be because the superposition of cross-domain information has no effect on the event 1 and event 2. As a result, it can effectively combine the virtues of different information sources to complement and enhance each other.



(a) On Flickr



(b) On Google News

Figure 7: The classification accuracy with the iteration of Gibbs sampling on Flickr data and Google News data, respectively.

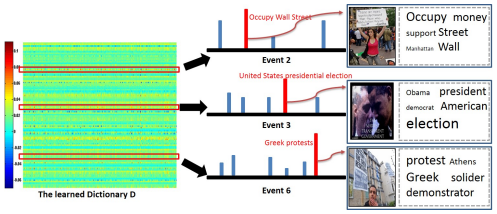


Figure 8: The visualization of the 100 learned dictionary elements. Each row corresponds to one element. The red color indicates the highest statistics results of binary vector z in each event set, and the corresponding dictionary element is shown with the image and text(All photos via Flickr under Creative Commons License).

In Figure 6, we give a detailed analysis about the proposed model. In Figure 6(a) and Figure 6(b), we show the comparison of the dictionary weights π_k (ordered in the probability) to be used by CDCL-s and CDCL on Flickr and Google News, respectively. We observe that most dictionary elements are used with a low probability in our CDCL, especially after the eighth dictionary element on Flickr. While the probability values of most dictionary elements used in CDCL-s are greater than 0.2. As a result, the learned feature representation of our CDCL is much more sparse. This shows that our model can effectively utilize the prior knowledge and experience of the auxiliary domain to learn a compact shared dictionary space and obtain the sparse representation. In Figure 6(c) and Figure 6(d), we show the statistics results of binary vector $\mathbf{z}$ by calculating the expected number of binary factors on Flickr and Google

News, respectively. We can see that the statistics results of binary vector $\mathbf{z}$ are consistent with the dictionary weights π_k, which shows that our model is reasonable. In Figure 7, we show the classification accuracies with the iteration of Gibbs sampling on Flickr and Google News, respectively. We can see that our CDCL can achieve accessible results after 20 iterations. As shown in Figure 8, we analyze the statistics results of binary vector $\mathbf{z}$ in each event and show the corresponding dictionary element (images and text) with the highest confidence score.

5.2 Cross-network Video Recommendation

5.2.1 Dataset Collection

For the cross-network video recommendation, we use the cross-network dataset [20]. This is a cross-network dataset with user account linkage between YouTube and Twitter, which contains 143,259 Google+ users, among which 38,540 users provide YouTube account, 39,400 users provide Twitter account, and 11,850 users provide both accounts. However, the dataset do not have the behavioral information of these users on Twitter. Therefore, we download the most recent $1,000$ tweets generated by each user via the official APIs according to the users ID provided by [20]. In order to better evaluate the recommendation results, we only use the users providing both the YouTube and Twitter accounts, and keep only the cross-network users who interacted with at least 8 different videos on YouTube. As a result, we obtain 1655 cross-network users and 5105 videos in total for this experiment evaluation.

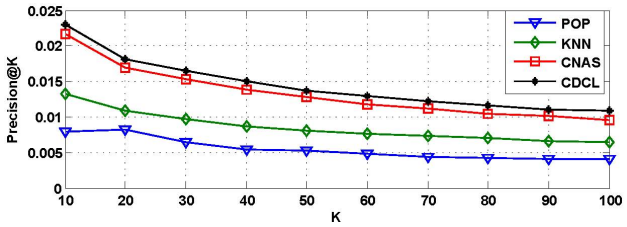
In our experiment setting, the proposed video recommendation solution is expected to facilitate cold-start recommendation for the new YouTube user. In the first stage of our cross-network video recommendation, we randomly select 900 active users to construct the training dataset, which is to learn the shared dictionary space. The remaining 755 users are considered as the cold-start users on YouTube, denoted as U^{new}. For the testing user $u_t \in U^{new}$, all the observed video-related interactions are hidden in the second stage and taken as ground truth for evaluation.

5.2.2 Evaluation Metrics

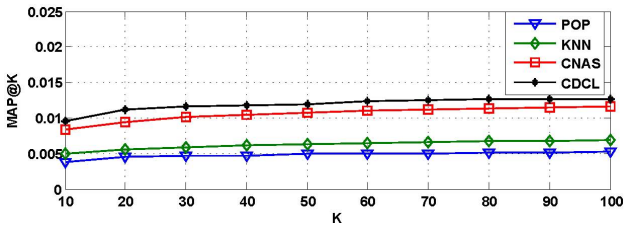
In practical video recommender system, users are basically only concerned about the top-ranked recommendation results and the available space to present the results on YouTube is also limited. The aim of the personalized video recommendation is to provide each user a ranking list of videos. Similar to traditional information retrieval task, we use Precision@K, Mean Average Precision (MAP@K) to measure the quality of the ranking list of recommended videos. For each new user $u \in U^{new}$ in the test set, Precision@K is defined as $Precision@K = \sum_{k=1}^{K} r_k/K$, and the MAP@K is the mean of average precision scores over test users U^{new} and is defined as:

$$MAP@K = \frac{1}{U^{new}} \sum_{u=1}^{U^{new}} \frac{\sum_{k=1}^{K} Precision@uk * r_{u,k}}{L_u}, \quad (14)$$

where r_k is the relevance level at position k, which is 0 for "Not Relevant" and 1 for "Relevant" in our experiment. The $r_{u,k}$ is the relevance level at position k for user u. The $Precision@uk$ is the precision at position k for user u, and K is the truncation level. L_u is the number of relevant data in the recommended set. The evaluation result is obtained by checking whether the recommended videos are truly in u's interested video set. In our study, we set $K \in 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$.



(a) Precision@K



(b) MAP@K

Figure 9: The Precision and MAP of cross-network video recommendation for new YouTube users.

5.2.3 Results and Analysis

We compare the proposed model with three baseline methods:

- Popularity(POP): This approach provides the same recommendation list of the videos to all new users according to this video's popularity. Here, we consider the view counts of the video as its popularity.
- KNN: For a new YouTube user, the KNN uses his/her Twitter information to obtain the most related Twitter users. Then the relevant videos can be obtained by the most related users.
- Cross-network Association (CNAS) [20]: The CNAS uses a coupled dictionary learning method to learn a pair of dictionary spaces by the same users across different networks.

The evaluation results of different methods are shown in Figure 9. From the results, we have the following conclusions: (1) The POP method shows inferior performance. This is due to its incapability of learning user's personalized needs and considering cross-network user behaviors. (2) The KNN and CNAS methods achieve better results. This shows that it is useful to adopt the auxiliary domain and consider the cross-network collaboration for the cold start recommendation task. (3) The proposed CDCL method outperforms the CNAS and the KNN, and achieves the best recommendation performance in terms of precision and MAP under all values of K computed. This is because the proposed model can collaboratively learn the shared dictionary space with the shared domain priors, which can better leverage users' cross-network activity data to address the user cold-start recommendation problem.

In Figure 10, we show four new YouTube users with their twitter history information on Twitter and the corresponding recommended video list from YouTube. Take the test user "Daniel Rodriguez" as an example, we can see that this new user is a software engineer and likes music and science. The corresponding recommended video list from YouTube includes some science technology, game design, and popular music, which better satisfies the interest of the new YouTube user. These results demonstrate the effectiveness of the proposed cross-domain collacross-network video recommendation method.

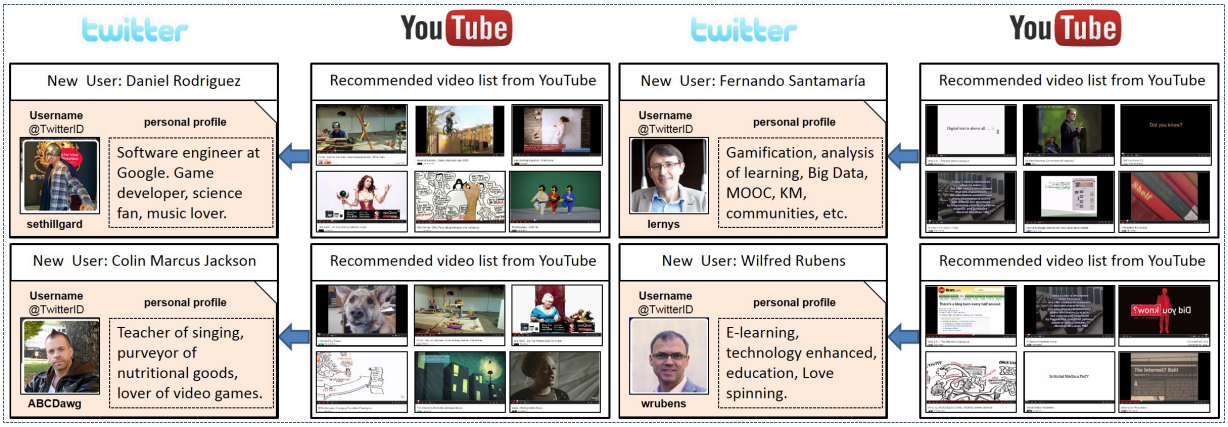


Figure 10: Four examples on cross-network video recommendation from Twitter to YouTube users(All photos via Flickr under Creative Commons License).

6. CONCLUSION

In this paper, we propose a generic cross-domain collaborative learning framework based on non-parametric Bayesian dictionary learning model for cross-domain data analysis. The proposed model can effectively adopt the shared domain priors and modality priors to collaboratively learn the feature representation by considering the domain discrepancy and the multi-modal property. The extensive experimental results on two different applications (cross-platform event recognition and cross-network video recommendation) demonstrate the effectiveness of the proposed model. In the future, we will investigate more applications with the proposed generic framework, such as cross-domain event summarization and cross-domain attribute mining.

7. ACKNOWLEDGMENT

This work is supported in part by National Basic Research Program of China (973 Program No. 2012CB316304), and National Natural Science Foundation of China (No.61225009, 61432019, 61303173, 61472115, 61472379, U1435211).

8. REFERENCES

[1] Xiaoshan Yang, Tianzhu Zhang, and Changsheng Xu. Cross-domain feature learning in multimedia. *IEEE Transactions on Multimedia*, 17(1):64–78, 2015.

[2] Tianzhu Zhang and Changsheng Xu. Cross-domain multi-event tracking via co-pmht. *ACM Trans. Multimedia Comput. Commun. Appl.*, 10(4):31:1–31:19, 2014.

[3] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain collaboration recommendation. In *KDD '12*, 2012.

[4] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *EMNLP '06*, 2006.

[5] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. In *IJCAI'09*, 2009.

[6] Fabian Abel, Samur Araújo, Qi Gao, and Geert-Jan Houben. Analyzing cross-system user modeling on the social web. In *Web Engineering - 11th International Conference, ICWE 2011, Paphos, Cyprus, June 20-24, 2011*, pages 28–43, 2011.

[7] Ming Yan, Jitao Sang, Tao Mei, and Changsheng Xu. Friend transfer: Cold-start friend recommendation with cross-platform transfer learning of social knowledge. In *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo, ICME 2013, San Jose, CA, USA, July 15-19, 2013*, pages 1–6, 2013.

[8] Xiaoshan Yang, Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Boosted multifeature learning for cross-domain transfer. *TOMCCAP*, 11(3):35:1–35:18, 2015.

[9] Yang Yang, Yi Yang, and Heng Tao Shen. Effective transfer tagging from image to video. *ACM Trans. Multimedia Comput. Commun. Appl.*, 9(2):14:1–14:20, 2013.

[10] Yang Yang, Zheng-Jun Zha, Yue Gao, Xiaofeng Zhu, and Tat-Seng Chua. Exploiting web images for semantic video indexing via robust sample-specific loss. *IEEE Transactions on Multimedia*, 16(6):1677–1689, 2014.

[11] Xiaoshan Yang, Tianzhu Zhang, Changsheng Xu, and M. Shamim Hossain. Automatic visual concept learning for social event understanding. *IEEE Transactions on Multimedia*, 17(3):346–358, 2015.

[12] Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, pages 216–228, 2015.

[13] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, and M. Shamim Hossain. Social event classification via boosted multimodal supervised latent dirichlet allocation. *TOMCCAP*, 11(2):27:1–27:22, 2014.

[14] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas S. Huang. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, 21(8):3467–3478, 2012.

[15] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *CVPR*, pages 2216–2223, 2012.

[16] Mingyuan Zhou, Haojun Chen, John William Paisley, Lu Ren, Guillermo Sapiro, and Lawrence Carin. Non-parametric bayesian dictionary learning for sparse image representations. In *NIPS'09*, pages 2295–2303, 2009.

[17] Chunfeng Yuan, Weiming Hu, Guodong Tian, Shuang Yang, and Haoran Wang. Multi-task sparse learning with beta process prior for action recognition. In *CVPR'13*, pages 423–429, 2013.

[18] Suman Deb Roy, Tao Mei, Wenjun Zeng, and Shipeng Li. Socialtransfer: cross-domain transfer learning from social streams for media applications. In *MM'12*, pages 649–658, 2012.

[19] Fabian Abel, Eelco Herder, Geert-Jan Houben, Nicola Henze, and Daniel Krause. Cross-system user modeling and personalization on the social web. *User Model. User-Adapt. Interact.*, 23(2-3):169–209, 2013.

[20] Ming Yan, Jitao Sang, and Changsheng Xu. Mining cross-network association for youtube video promotion. In *MM '14*, 2014.

[21] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31:210–227, 2009.

[22] John Paisley and Lawrence Carin. Nonparametric factor analysis with beta process priors. In *ICML '09*, pages 777–784, 2009.

[23] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[24] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[25] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

[26] Timo Reuter, Symeon Papadopoulos, Georgios Petkos, Vasileios Mezaris, Yiannis Kompatsiaris, Philipp Cimiano, Christopher M. De Vries, and Shlomo Geva. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In *MediaEval*, 2013.

[27] L. Liu, L. Wang, and X. Liu. In defense of softassignment coding, 2011. In ICCV.

[28] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *MM '10*, pages 251–260, 2010.