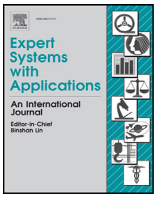
Expert Systems With Applications 104 (2018) 168–184

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



MMDF-LDA: An improved Multi-Modal Latent Dirichlet Allocation model for social image annotation

Liu Zheng[a,b,*], Zhang Caiming[b,c,d,e], Chen Caixian[a]

[a] School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China
[b] Shandong Provincial Key Laboratory of Digital Media Technology, Shandong University of Finance and Economics, Jinan 250014, China
[c] School of Computer Science and Technology, Shandong University, Jinan 250014, China
[d] Shandong Co-Innovation Center of Future Intelligent Computing, Yantai 264025, China
[e] School of Information and Electrical Engineering, Ludong University, Yantai 264025, China

ARTICLE INFO

Article history:
Received 19 November 2017
Revised 7 February 2018
Accepted 10 March 2018
Available online 12 March 2018

Keyword:
Social image
Multi-modal data fusion
LDA model
Semantic annotation
Geographical topic

ABSTRACT

Social image annotation, which aims at inferring a set of semantic concepts for a social image, is an effective and straightforward way to facilitate social image search. Conventional approaches mainly demonstrated on adopting the visual features and tags, without considering other types of metadata. How to enhance the accuracy of social image annotation by fully exploiting multi-modal features is still an opening and challenging problem. In this paper, we propose an improved Multi-Modal Data Fusion based Latent Dirichlet Allocation (LDA) topic model (MMDF-LDA) to annotate social images via fusing visual content, user-supplied tags, user comments, and geographic information. When MMDF-LDA samples annotations for one data modality, all the other data modalities are exploited. In MMDF-LDA, geographical topics are generated from GPS locations of social images, and annotations have different probability to be used in different geographical regions. A social image is divided into several patches in advance, and then MMDF-LDA assigns annotations for the patches of social images by estimating the probability of annotation-patch assignment. Through experiments in social image annotation and retrieval on several datasets, we demonstrate the effectiveness of the proposed MMDF-LDA model in comparison with state-of-the-art methods.

1. Introduction

In the era of Web 2.0, the rising prominence of photo sharing websites (e.g. Flickr (Flickr)) has led to a great explosion of social images (Xie, Yu, & Hu, 2014). Some typical Web 2.0 platforms, such as Facebook, Twitter and Flickr, allow users not only to upload and share their multimedia data with other users, but also to provide semantic descriptive terms (Li & Tang, 2017). Social media sharing websites allow users to annotate images with free tags, which greatly contributes to the development of image retrieval (Cheng, Shen, & Miao, 2016). Social media sharing websites provide many challenging and interesting issues for us; for instance, how to fully utilize various types of user-supplied contextual information.

As the number of user-supplied social images is dramatically increasing, there is a growing requirement to search relevant social images for users, and social image analysis and retrieval are important to help people organize and access the increasing amount of user-tagged multimedia (Li, Liu, Tang, & Lu, 2014; Li, Snoek, & Worring, 2008; Li, Snoek, & Worring, 2009). Hence, effectively retrieving social images in the large scale database with high accuracy is of great importance. From the beginning in the 90s of the last century, content-based image retrieval (CBIR) systems have been studied and developed by many researchers (Carson, Belongie, Greenspan, & Malik, 2002; Ireton, & Xydeas, 1991; Jing, & Baluja, 2008; Ma, & Manjunath, 1999; Pentland, Picard, & Sclaroff, 1996; Smeulders, Worring, Santini, Gupta, & Jain, 2000). CBIR system retrieves images by calculating the similarity between image visual features. However, performance of the existing CBIR systems has not been satisfied by us (Cheng et al., 2016). The reason lies in that the semantic gap between high level semantic information and low level visual features still cannot be tackled well. Different from traditional CBIR systems, Web 2.0 platform allows users to share, evaluate and upload multimedia information, as well as to provide the social images with semantic terms (tags) and other contextual information (Lu, Liu, & Qian, 2016). Therefore, Web 2.0 based platform provides us new opportunities to enhance the accuracy of social image annotations.

[*] Corresponding author.
E-mail addresses: lzh_48@126.com (L. Zheng), czhang@sdu.edu.cn (Z. Caiming), ccxsdufe@sina.com (C. Caixian).

https://doi.org/10.1016/j.eswa.2018.03.014
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

As the CBIR system cannot retrieve images effectively, researchers have designed and developed the text-based image retrieval system (TBIR), which retrieves images by estimating the similarity between query words and manual annotations. However, manually annotating social images is quite time consuming (Liu, Yuan, Cong, & Xu, 2014; Xia, Peng, Feng, & Fan, 2014). Hence, it is important to mine the semantic information from social images and provide them several descriptive words. Additionally, famous social images sharing websites (such as Flickr and Photosig) host a huge number of social images with user-supplied tags (Lei, Liu, & Li, 2016). Unfortunately, not all user-supplied tags are actually relevant to the image visual content, because the tagging behavior is thoroughly free without any rules and constraints. For example, some users assign the camera model and personal information to images, which are meaningless to describe the image semantics (Qian, Hua, Tang, & Mei, 2014).

Currently, most existing image tagging methods mainly concentrate on mapping visual features to semantic words, however, it limits the annotation accuracy. Unlike the traditional methods, in social image sharing websites, images are often accompanied by different types of metadata like visual features, tags, ratings, comments, EXIF, as well as with the personal information about the uploaders and their own social relations. On one hand, multi-type of metadata make the image annotation task more difficult and complex; on the other hand, it also provides great opportunities for us to deeply study. In this paper, we try to annotate social images through effectively fusing different types of data based on a multi-modal LDA topic model, which is a generative probabilistic model of a corpus. In the LDA model, documents are organized as random mixtures on latent topics, and topics are represented as word distributions.

Compared with existing works, main contributions of our study are summarized as follows.

1) We present an improved Multi-Modal Latent Dirichlet Allocation model (MMDF-LDA) to analyze and mine the relations between input multi-modal data patterns and output semantic descriptions by mapping a high dimensional vector space to a low dimensional vector space. In MMDF-LDA, multi-modal data are described in a unified probabilistic topic space.
2) The MMDF-LDA model can construct an effective semantic description space and narrow the "semantic gap" by fully integrating multiple types of data, such as visual content, user-supplied tags, user comments, and geographic information. Furthermore, the MMDF-LDA model can easily be extended with other types of data modalities and also permit data missing.
3) The MMDF-LDA model provides annotations for each patch of social images (patch level social image annotation). When we sample annotations for image patches, all the other data modalities are exploited. Compared with annotating the whole image, MMDF-LDA provides fine-grained semantic representations for social images.
4) The geographical topic is defined for MMDF-LDA, and each geographical region is corresponding to a geographical topic distribution. Particularly, geographical topic may help us to find own semantic information and vision pattern of each region according to the geographical topic distribution.
5) Three different types of Bag of Words (BoW) model based visual feature descriptions are utilized in MMDF-LDA.

The rest of the paper is organized as follows. Section 2 presents a brief review of some relevant works. In Section 3, we state the social image annotation problem. In Section 4, a multi-modal data fusion model is developed to assign annotations for social images. Our experimental results on algorithm evaluation are provided in Section 5, and we conclude the whole paper and point the future works in Section 6.

2. Literature review

In the following, we briefly review those methods which are the most relevant to our research along two directions: (1) Applications of the topic model, and (2) Automatic image annotation.

2.1. Applications of the topic model

To effectively fuse multi-modal data, we introduce the topic model to solve the social image annotation issue. In the past few years, topic model has gained more and more attention from both academia and industry. A topic is composed of a cluster of words frequently co-occurred, that is, given a corpus, a topic model is able to distinguish words with various semantic meanings and then obtains hidden topics.

LDA model has been widely used in intelligent computing over the past decade. Blei et al. initially proposed the original LDA utilizing EM estimation (Blei, Ng, & Jordan, 2003). Griffiths and Steyvers exploited Gibbs sampling to predict parameters in the LDA model (Griffiths, & Steyvers, 2004). There are several researchers who have used LDA in data dimension reducing (Blei et al., 2003; Griffiths, & Steyvers, 2004; Quelhas et al., 2005), and very positive experimental results have been achieved.

LDA can also exploit the prior knowledge to enhance the algorithm performance. Andrzejewski and Zhu et al. proposed the z-label constraints as a mechanical modification to the Gibbs sampling equation (Andrzejewski & Zhu, 2009). Andrzejewski et al. utilized domain knowledge via a Dirichlet Forest prior in a LDA model (Andrzejewski, Zhu, & Craven, 2009). Hu et al. proposed an interactive topic model, which allows users to iteratively refine the topics using constraints to enforce some sets of words within a same topic (Hu, Boyd-Graber, Satinoff, & Smith, 2014). Zhai et al. presented a soft constraint named probabilistic must-link and cannot-link constraints, which refer to a relaxation mechanical modification in the Gibbs sampling equation (Zhai, Liu, Xu, & Jia, 2011). He et al. used a nonlinear Compressed Sensing-Based LDA Topic Model for SAR image classification (He, Zhuo, Ou, Liu, & Liao, 2014). Makita et al. exploited LDA-Based topic model to label blog posts using Wikipedia entries (Makita et al., 2013). Li et al. presented the Tagger Tag Resource-LDA-Community model (TTR-LDA-Community model) to integrate LDA model with the Girvan-Newman community discovery algorithm via an inference mechanism (Li et al., 2011). Lu et al. compared the probabilistic latent semantic analysis (PLSA) model with the LDA topic model in the domain of information retrieval using three typical tasks, such as document clustering, text categorization, and ad-hoc retrieval (Lu, Mei, & Zhai, 2011).

Different from the above works, Virtanen et al. proposed the a Factorized Multi-Modal Topic Model (FMMTM), which integrates two approaches by presenting a novel hierarchical Dirichlet process based topic model that automatically learns both shared and private topics. In addition, this model is able to query the content of one domain given samples of the other (Virtanen, Jia, Klami, & Darrell, 2012). The fundamental assumptions of FMMTM and our proposed model are different. FMMTM supposes that different modalities have weak relationships, while our model supposes that different modalities have strong relationships, which conforms to the characteristics of social images.

2.2. Automatic image annotation

Automatic image annotation aims at assigning a set of semantic labels to images with the goal of bridging the so-called semantic gap between the available image features and the words which people might exploit to annotate images. Automatic image annotation has been extensively studied in recent years (Gupta, Li, Yin, & Han, 2010; Hanbury, 2008; Li et al., 2016; Tousch, Herbin, & Audibert, 2012; Zhang, Islam, & Lu, 2012), and some typical works are presented as follows.

We firstly discuss the earlier methods about image annotation. Liu et al. proposed an automatic image annotation approach based on an adaptive similarity graph model and the manifold ranking learning, in which the visual and textual information are integrated (Liu, Li, Ma, Liu, & Lu, 2006). Following the basic hypothesis that annotations provide evidence for the class label, and the class label provides evidence for annotations. Chong et al. proposed a novel probabilistic model for jointly modeling the image, its class label, and its annotations based on an approximate inference and estimation algorithm (Chong, Blei, & Li, 2009). Wang et al. aimed to reduce the semantic gap in Web image retrieval and annotation. Main contributions of this work lie in that a ranking-based distance metric learning method and a LDA based semantic similarity computing method are proposed (Wang, Zhang, & Zhang, 2008). In addition, other models or technologies were used in image annotation before 2010, such as Multi-label sparse coding (Wang, Shuicheng, Lei, & Zhang, 2009), Markov chain (Wang, Jing, Zhang, & Zhang, 2007), Graph learning (Liu, Li, Liu, Lu, & Ma, 2009), Ontologies (Srikanth, Varner, Bowden, & Moldovan, 2005), and Probabilistic semantic model (Zhang, Zhang, Li, Ma, & Zhang, 2005).

Recently, image annotation is still an important and hot issue, and many researchers have devoted themselves to solve it.

Supervised, unsupervised and semi-supervised learning have been utilized in image annotation. Pellegrin et al. presented two effective methods for unsupervised automatic image annotation in the context of a common framework, which is inspired in the way that a query is expanded throughout Automatic Query Expansion in information retrieval (Pellegrin, Escalante, Montes-y-Gómez, & González, 2017). Applying both labeled images and unlabeled images to uncover the intrinsic data structural information, Hu et al. proposed an effective and robust scheme to facilitate the image annotation task (Hu et al., 2017). Song et al. proposed a semi-supervised image annotation method via learning an optimized graph from multi-cues, which can more accurately embed the relationships among the data points (Song et al., 2016).

Tensor or matrix based models also can be used in image annotation. Recently, Tang et al. proposed a tri-clustered tensor completion framework to collaboratively explore the visual, tag and user information to improve the performance of social image tag refinement. Particularly, the inter-relations among users, images and tags are represented as a tensor, and the intra-relations between users, images and tags are explored by three regularizations (Tang et al., 2017). To obtain high-quality tags from user-supplied tags with missing information and noise, Zhang et al. proposed a unified tag matrix completion framework by learning the image-tag relation within each Points of Interests (POI). In particular, the model optimization are parallelized and distributed to learn the tag submatrix for each POI (Zhang, Wang, & Huang, 2017).

Other algorithms or theories in machine learning have been exploited to handle image annotation as well. Verma et al. proposed a 2-pass k-nearest neighbor algorithm by integrating both the image-to-label similarity and the image-to-image similarity. Furthermore, a metric learning framework is proposed as well (Verma & Jawahar, 2017). Jiu et al. plugged a multiple kernel learning model into support vector machines for image annotation. In addition, four different frameworks are provided to learn the weights of these networks, such as supervised, unsupervised, kernel-based semi-supervised, and Laplacian-based semi-supervised (Jiu & Sahbi, 2017). Wu et al. proposed a novel partially labeled factor graph model (also named as the demographics factor graph model) to infer emotional tags from social images (Wu et al., 2017). Bahrololoum et al. developed a multi-expert based framework for automatic image annotation based on the integration of results which are obtained from both feature space and concept space (Bahrololoum & Nezamabadi-pour, 2017).

With the rapid development of online social networks, social computing has provided us new ideas to annotate images. Lei et al. presented a social diffusion analysis method for image annotation, which utilizes massive social diffusion records about how images are disseminated within online social networks. In this work, user preferences are represented as common interests of pairwise users instead of individual user interest (Lei et al., 2016). Gu et al. proposed an image annotation method which utilizes latent semantic community of labels and multiple kernel learning, and a candidate label ranking based method is determined by intracommunity and intercommunity ranking (Gu et al., 2015).

Apart from the above works, other ideas and models have been adopted in image annotation, such as Multi-Instance Multi-Label Learning (Ding et al., 2016), Multi-Label Dictionary Learning (Jing, Wu, Li, Hu, & Zhang, 2016), Fuzzy-knowledge representation (Ivasic-Kos, Pobar, & Ribaric, 2016), Semantic Concept Co-Occurrence (Feng & Bhanu, 2016), and Semantic Label Embedding Dictionary Representation (Cao, Zhang, Guo, Liu, & Meng, 2015).

Social image annotation is a typical multi-modal data fusion problem, which has been utilized in several applications over the past decade. Multi-modal data fusion has gained researchers' wide attention, because it can effectively solve different multimedia analysis tasks. Multi-modal data fusion is defined as integrating multiple media with their associated features to perform the data analysis and mining task (Atrey, Hossain, Saddik, & Kankanhalli, 2010). In the following, we will introduce related works about multi-modal data fusion.

Chavali et al. presented a sequential and hierarchical Monte Carlo Bayesian system for state prediction utilizing multi-modal data (Chavali & Nehorai, 2014). Berger et al. proposed a novel multi-modal and multi-temporal data fusion method based on the data in the 2012 GRSS Data Fusion Contest (Berger et al., 2013). Soumya et al. proposed a comprehensive multi-modal structural analysis process which contains intra-and inter-modal nondestructive evaluation data fusion based on eddy current, millimeter wave, and ultrasonic data obtained from 5 lap-joint mimic test panels (De et al., 2013). Longbotham et al. proposed four awarded algorithms and the conclusions of the contest, investigated both supervised and unsupervised methods, and then exploited multi-modal data to detect flood disaster (Longbotham et al., 2012). Noore et al. proposed a multi-level wavelet based fusion algorithm integrating fingerprint, face, iris, and signature images together (Noore, Singh, & Vatsa, 2007). In addition, multi-modal data fusion has also been used in face detection (Chen, Flynn, & Bowyer, 2006) and hierarchical clustering algorithm to group mixed type data (Coppock & Mazlack, 2004).

In addition, LDA model and its modified versions have been used in image annotation.

Lienou et al. used LDA to annotate satellite images by classifying patches of the large images and integrating the spatial information between these patches (Lienou, Maitre, & Datcu, 2010). Following the assumption that images and their co-occurring textual data are generated by mixtures of latent topics, Feng et al. proposed a probabilistic model to annotate images by exploiting the vast resource of images and documents available on the Web (Feng & Lapata, 2010). Xu et al. proposed a regularized Latent Dirichlet Allocation (rLDA) model to annotate images by exploiting both the statistics of tags and visual affinities of images in the corpus. Furthermore, tag similarity and tag relevance are jointly estimated in an iterative manner (Xu, Wang, Hua, & Li, 2009). Putthividhya et al. proposed a new probabilistic model (sLDA-bin) for the task of image annotation. sLDA-bin is extended from supervised Latent Dirichlet Allocation (sLDA) model. Different from the correspondence LDA (cLDA), the association model in sLDA permits each caption word to be associated with more than one image region, hence, it is more appropriate for annotations which globally describe the scene (Putthividhya, Attias, & Nagarajan, 2010a).

Multi-modal data fusion based LDA models have been exploited to annotate image, and these models are similar to our proposed model. As a modified version of LDA, Corr-LDA (Blei & Jordan, 2003) is able to learn the joint distribution of different types of data. To mine correlations between two data modalities, the association models utilize a set of shared latent variables to represent the underlying causes of cross-correlations in the data. Tr-mmLDA (Putthividhya, Attias, & Nagarajan, 2010b) refers to a statistical topic model for multimedia data (e.g. images and videos) annotation. The main innovations of this method lie in that correlations between image or video features and annotations are captured by a latent variable regression method. Unlike sharing a set of latent topics between two data modalities as in the formulation of Corr-LDA, Tr-mmLDA develops a regression module to correlate two sets of topics. M3LDA (Nguyen, Zhan, & Zhou, 2013) is another LDA topic model based image annotation approach, which is made up of (1) Visual-label part, (2) Textual-label part and (3) Label-topic part. Combining visual features and user-supplied tags together, labels are assigned to images.

The main difference with the proposed approach and the above LDA based image annotation methods mainly lies in that the topics in MMDF-LDA are generated from geographic regions instead of images.

As compared to the above related works, the key contributions of this paper are summarized towards the end of Section 1.

3. Statement of the social images annotation problem

Social image annotation is becoming increasingly popular with the rapid development of the Web 2.0 platform, in which images are annotated with arbitrary tags by users. Most of the existing image annotating methods mainly concentrate on analyzing visual similarity or mapping visual features to tags. Social images own multi-modal data, such as visual features, user supplied tags, and persons' comments. Therefore, the performance of social images annotating highly relies on the performance of multi-modal data fusion. Formal description of the social image is listed as follows.

Definition 1 (**Social image**)**.** A social image is provided by a common user on photo sharing website such as Flickr, Instagram, and SmugMug. Different from traditional images, social images have rich metadata, including user-supplied tags, user contact information, geographic information, EXIF, etc.

Different from traditional text annotation system, social image annotation system must solve several difficult problems, such as, semantic gap between the lower level visual features and the higher level semantic information. In particular, there is a huge amount of noisy information in social media, because social media are freely provided by users. In the social image sharing websites, images are associated with rich textual information, e.g. user supplied tags. On the other hand, other types of data (e.g. author contact information, geographic information and user group) can help us to mine semantic information as well. Hence, effectively integrating the multi-modal attributes of the social images can significantly promote the performance of information retrieval and data mining under the social media environment.

Traditional image annotation methods aim to assign appropriate semantic terms for a given image (denoted as I) from a fixed annotation vocabulary (denoted as A). This process is formally described as follows.

(1) Construct a training image dataset $TR=\{TR_1,TR_2,\cdots, TR_N\}$, where TR_i refers to an annotated image. We suppose that TR_i has been manually annotated by a semantic term list. Moreover, TR_i is represented as a set of visual feature vectors $F_i = \{f_i^1, f_i^2, \cdots, f_i^{Q_i}\}$, where f_i^j is the j^{th} visual feature vector and Q_i denotes the number of visual feature vectors of TR_i.
(2) The annotation vocabulary A is represented as $A=\{a_1,a_2,\cdots, a_{|A|}\}$.
(3) Annotations of image I are represented as $A_I = \{a_{I1}, a_{I2}, \cdots, a_{I|A_I|}\}$.

Each training image TR_i contains a set of images with annotations. For an unlabeled testing image I, the probability of word a to be annotated for it is estimated as follows.

$$p(a|I) \propto p(a, I) = \sum_{i=1}^{|TR|} p(I|TR_i) \cdot p(a|TR_i) \cdot p(TR_i) \qquad (1)$$

where $p(TR)$ follows the uniformly distribution, and $p(I|TR_i)$ refers to the probability of I generated from TR_i. Additionally, $p(a|TR_i)$ is higher when the image TR_i is annotated by word a. For the testing image I, all words in A are ranked according to Eq. (1), and the top ranked ones are reserved as the final annotations (A_I).

An example of a social image in Flickr with multi-modal metadata is illustrated in Fig. 1.

Fig. 1 describes an example of a social image[1] in Flickr. Additionally, this example is titled "still life", and it describes a lonely bay sunrise in New Zealand. Different from the traditional Web images, social images have various types of Metadata, such as author, title, user comments, user supplied tags. In particular, fully utilizing all these multi-modal metadata can greatly enhance the accuracy of social image annotations.

To take full advantage of different types of metadata in social images annotation, we discuss how to use the probabilistic topic model to fuse multi-modal metadata. That is, we aim to effectively mine the potential association between the unlabeled images and the labeled images, and then learn a probability model to automatically annotate new images by fusing multi-modal metadata.

4. The proposed social image annotation method

In this section, we propose the MMDF-LDA model to provide annotations for social images by fusing the multi-modal metadata of social images.

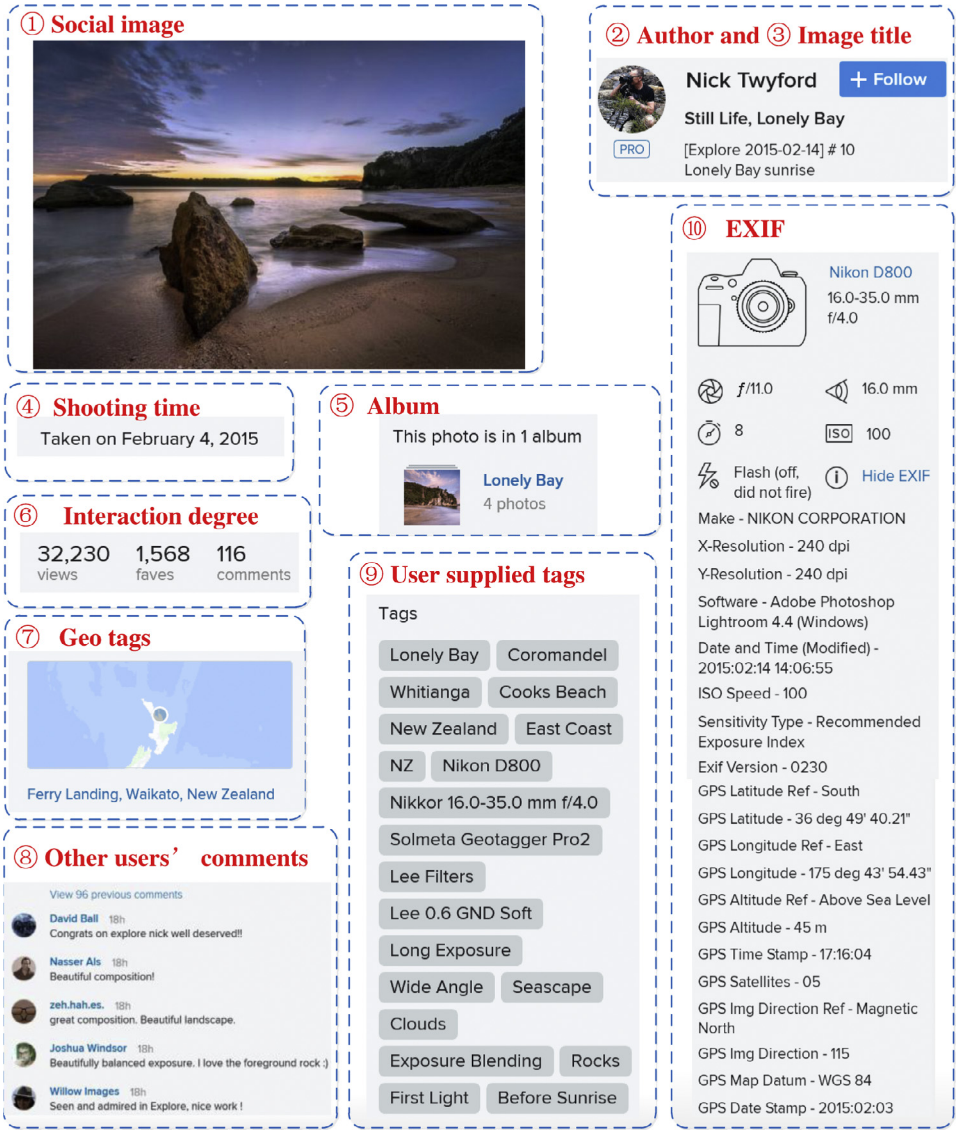
4.1. Preliminary for LDA topic model

To solve the issue of sharing latent variables between different data modalities, in this paper, we extend the LDA topic model to effectively model correlations between various types of metadata.

LDA refers to a generative probabilistic model of a corpus. In the LDA model, documents are represented as random mixtures of latent topics, and each topic is represented as a word distribution (Blei et al., 2003). A word is defined as an element from a vocabulary, and a document refers to a sequence of NW words. Moreover, a corpus denotes a collection of MD documents. Framework of the LDA model is illustrated in Fig. 2.

In Fig. 2, LDA is represented as a probabilistic graphical model with three levels, in which the outer plate represents documents, and the inner plate refers to the repeated choice of topics and

[1] https://www.flickr.com/photos/67654596@N04/16522119302/in/explore-2015-02-14.



Nick Twyford
Still Life, Lonely Bay
[Explore 2015-02-14] # 10
Lonely Bay sunrise

Taken on February 4, 2015

This photo is in 1 album
Lonely Bay
4 photos

32,230 views 1,568 faves 116 comments

Ferry Landing, Waikato, New Zealand

View 96 previous comments

David Ball 18h
Congrats on explore nick well deserved!!

Nasser Als 18h
Beautiful composition!

zeh.hah.es. 18h
great composition. Beautiful landscape.

Joshua Windsor 18h
Beautifully balanced exposure. I love the foreground rock :)

Willow Images 18h
Seen and admired in Explore, nice work !

Tags
Lonely Bay, Coromandel, Whitianga, Cooks Beach, New Zealand, East Coast, NZ, Nikon D800, Nikkor 16.0-35.0 mm f/4.0, Solmeta Geotagger Pro2, Lee Filters, Lee 0.6 GND Soft, Long Exposure, Wide Angle, Seascape, Clouds, Exposure Blending, Rocks, First Light, Before Sunrise

Nikon D800
16.0-35.0 mm
f/4.0
ƒ/11.0
16.0 mm
8
ISO 100
Flash (off, did not fire)
Make - NIKON CORPORATION
X-Resolution - 240 dpi
Y-Resolution - 240 dpi
Software - Adobe Photoshop Lightroom 4.4 (Windows)
Date and Time (Modified) - 2015:02:14 14:06:55
ISO Speed - 100
Sensitivity Type - Recommended Exposure Index
Exif Version - 0230
GPS Latitude Ref - South
GPS Latitude - 36 deg 49' 40.21"
GPS Longitude Ref - East
GPS Longitude - 175 deg 43' 54.43"
GPS Altitude Ref - Above Sea Level
GPS Altitude - 45 m
GPS Time Stamp - 17:16:04
GPS Satellites - 05
GPS Img Direction Ref - Magnetic North
GPS Img Direction - 115
GPS Map Datum - WGS 84
GPS Date Stamp - 2015:02:03

Fig. 1. An example of a social image in Flickr with rich metadata.

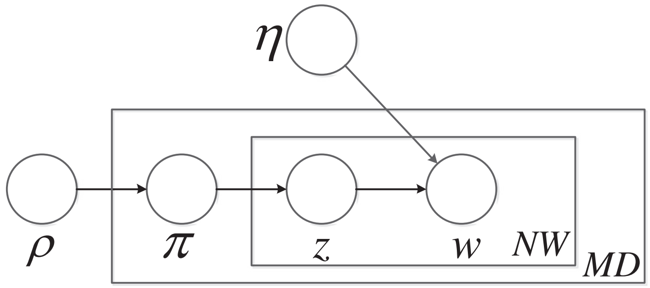


Fig. 2. Framework of the LDA model.

words in a document. Documents are associated with multiple topics in the LDA model, and the generative process for each document w in a corpus CP is represented as follows (Blei et al., 2003).

1) Choose $NW \sim Poisson(\xi)$
2) Choose $\pi \sim Dir(\rho)$
3) For each word w_n, $n \in [1, NW]$
   i) Choose a topic $z_n \sim Mul(\pi)$
   ii) Choose a word w_n from $p(w_n|z_n,\eta)$, which denotes a multinomial probability conditioned on the topic z_n.

Some assumptions are provided for LDA, including: 1) the dimensionality of the Dirichlet distribution is fixed, and 2) the word probabilities are parameterized by a matrix η, where $\eta_{ij} = p(w^j = 1|z^i = 1)$. Moreover, NW is not dependent of all the other data generating variables, such as π and z.

Given parameters ρ and η, the joint distribution of a topic mixture π, a set of NW topics $\mathbf{z}$, and a set of NW words $\mathbf{w}$ is estimated as follows.

$$p(\pi, \mathbf{z}, \mathbf{w}|\rho, \eta) = p(\pi|\rho) \prod_{n=1}^{NW} p(z_n|\pi) p(w_n|z_n, \eta) \tag{2}$$

Afterwards, the probability of a corpus can be computed using the product of the marginal probabilities of a single document as follows.

$$p(CP|\rho, \eta) = \prod_{d=1}^{MD} \int p(\pi_d|\rho) \times \left(\prod_{n=1}^{NW_d} \sum_{z_{dn}} p(z_{dn}|\pi_d) p(w_{dn}|z_{dn}, \eta) \right) d\pi_d \tag{3}$$

Poisson and Dirichlet distributions are important theoretical basis for LDA, and their definitions are given as follows.

Definition 2 (**Poisson distribution**)**.** A discrete random variable X is said to have a Poisson distribution with parameter $\chi > 0$, if, for $m = 0, 1, 2, \cdots$, the probability mass function of X is given by:

$$f(m; \chi) = \Pr(X = m) = \frac{\chi^m e^{-\chi}}{m!} \tag{4}$$

where e is the Euler's number and $m!$ is the factorial of m.

Definition 3 (**Dirichlet distribution**)**.** The Dirichlet distribution of order $U \geq 2$ with parameters $\alpha_1, \alpha_2, \cdots, \alpha_U > 0$ has a probability density function with respect to Lebesgue measure on the Euclidean space R^{U-1} given by:

$$f(x_1, x_2, \cdots, x_U; \alpha_1, \alpha_2, \cdots, \alpha_U) = \frac{1}{B(\alpha)} \prod_{u=1}^{U} x_u^{\alpha_u - 1} \tag{5}$$

where $\{x_u\}_{u=1}^{u=U}$ belongs to the standard $U-1$ simplex, or in other words:

$$\sum_{u=1}^{U} x_u = 1 \quad \text{and} \quad x_u \geq 0 \quad \text{for all} \quad u \in [1, U] \tag{6}$$

The normalizing constant is the multivariate Beta function, which is represented based on the gamma function:

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \ \alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K) \tag{7}$$

where $\Gamma()$ is the gamma function.

4.2. Visual features extraction

Visual feature is a fundamental and important modal in social image annotation. As LDA uses the BoW model to represent data, in this work, three different types of BoW based visual feature descriptions are utilized, such as 1) BoVW (Bags of Visual Word)+SIFT, 2) DBoVW (Descriptive Bags of Visual Word)+SIFT, and 3) BoVW+SIFT+CNN (Convolutional Neural Network). In particular, we partition the image I to $m \times n$ patches ($I = (pt_1, pt_2, \cdots, pt_{m \times n})$), where pt_i refers to the ith patch of I.

4.2.1. BoVW+SIFT

Following Li's work (Li & Perona, 2005), visual content of social images are described as 128-dim scale invariant feature transform (SIFT) descriptors computed on several gray-scale patches. Furthermore, 36-dim robust color descriptors are used to complement the SIFT descriptors. We sample a fixed number of keypoints per image, and all SIFT descriptors together with color descriptors of these keypoints are clustered to create a "visual vocabulary" (also named as codebook). Afterwards, K-means algorithm is executed on a set of 164-dim features to learn a visual dictionary with 256 visual words. Each image patch is represented as a 256-dim feature (BoVW+SIFT) vector which indicates how many times each visual word occurs in an image patch (except the NUS-WIDE dataset). To represent different statistics in different image datasets, visual word dictionary should be learned separately for each image dataset.

4.2.2. DBoVW+SIFT

Currently, visual words are usually obtained by unsupervised clustering methods which may generate many unnecessary and non-descriptive words with a low discriminative power. Hence, how to make visual words be descriptive is of great importance for computer vision and image retrieval (Liu, Zhang, Zhuo, & Yang, 2015). Following Zhang et al.'s work (Zhang, Yang, Tian, Zhuo, & Liu, 2017), we only select the visual words which have the high discriminative power to represent the image features by a visual word ranking algorithm, and then construct the DBoVW model. For a particular image category C_t, we construct a matrix $R(C_t)$ to integrate the frequency with co-occurrence of visual words. The diagonal element $R_{i,i}^{C_t}$ of $R(C_t)$ represents the inherent importance of category C_t, on the other hand, the non-diagonal element $R_{i,j}^{C_t}$ denotes the average co-occurrence frequency of visual word i and j. Afterwards, just like the famous Pagerank algorithm (Page, Brin, Motwani, & Winograd, 1999), the initial rank score of each visual word is equal, and then we start the ranking scores updating iterations. The most descriptive visual words are obtained via choosing the top ranked visual words (Liu et al., 2015; Zhang et al., 2017).

For a given image patch, we exploit the traditional BoVW features, and only reserve the dimensions which are related to the top ranked visual words. Accordingly, values of other dimensions are all set to zero. Particularly, the DBoVW model is also constructed using the SIFT descriptor in this work. Furthermore, the visual dictionary with 256 visual words is learned (except the NUS-WIDE dataset), and each image patch is represented as a 256-dim feature (DBoVW+SIFT) vector, which is just the same as BoVW+SIFT.

4.2.3. BoVW+SIFT+CNN

As the CNN has been successfully used in many areas, such as Speech Recognition (Abdel-Hamid et al., 2014), and Human Action Recognition (Ji, Xu, Yang, & Yu, 2013). Moreover, Yan et al. have proved that the complementarity between CNN and SIFT features for image retrieval (Yan, Wang, Liang, Huang, & Tian, 2016). Therefore, we follow the works of Zheng, Wang, Wang, & Tian (2016) and Yan et al. (2016) to exploit the CNN descriptors to provide semantic information for SIFT descriptors. Afterwards, CNN descriptors and SIFT descriptors are fused together.

SIFT descriptors extracted from a particular image patch are aggregated using VLAD (vector of locally aggregated descriptors) (Jegou et al., 2012). It means that a set of SIFT descriptors is represented as a single fixed-size vector by VLAD.

For each image dataset, a codebook $CB = (cb_1, cb_2, \cdots cb_k)$ of k visual words is learned by K-means, and each visual word denotes the centroid of a cluster. Vector s of a SIFT descriptor is assigned to its nearest visual word $cb_i = NN(s)$.

$$NN(s) = \arg\min_{cb_i} \|s - cb_i\|, \ i \in \{1, 2, \cdots, k\} \tag{8}$$

For each visual word cb_i, the differences $s - cb_i$ of the vector s which is assigned to cb_i are accumulated as follows.

$$f_{v_i} = \sum_{s \text{ such that } NN(s) = cb_i} s - cb_i \tag{9}$$

Afterwards, a $128 \times k$-dim vector f_v is constructed as follows.

$$f_v = [f_{v_1}, f_{v_2}, \cdots, f_{v_i}, \cdots, f_{v_k}], \ f_{v_i} \in \mathbb{R}^{128} \tag{10}$$

Then, vector f_v is L_2-normalized as $f_v = f_v / \|f_v\|_2$, and f_v is named as the point-level feature. In particular, we set k=256, and then apply the PCA algorithm to reduce the dimension of f_v from 32,768 (128×256) to 1024.

Next, we discuss how to extract CNN features from image patches, therefore, in this work, CNN features are named as Patch-level features, which are used to capture the high level semantic information for an image by CNN. It is well known that the deep learning features, especially those from the high layers, are suitable to represent semantic information (Yan et al., 2016). Following the work of Yan et al. (2016), we extract the CNN features (f_p) of a specific image patch from the pool5 layer in GoogLeNet (Szegedy et al., 2015), which is a famous deep convolutional network, and ranked as the top one at the image classification task in ILSVRC 2014. We set the dimension of f_p as 1024.

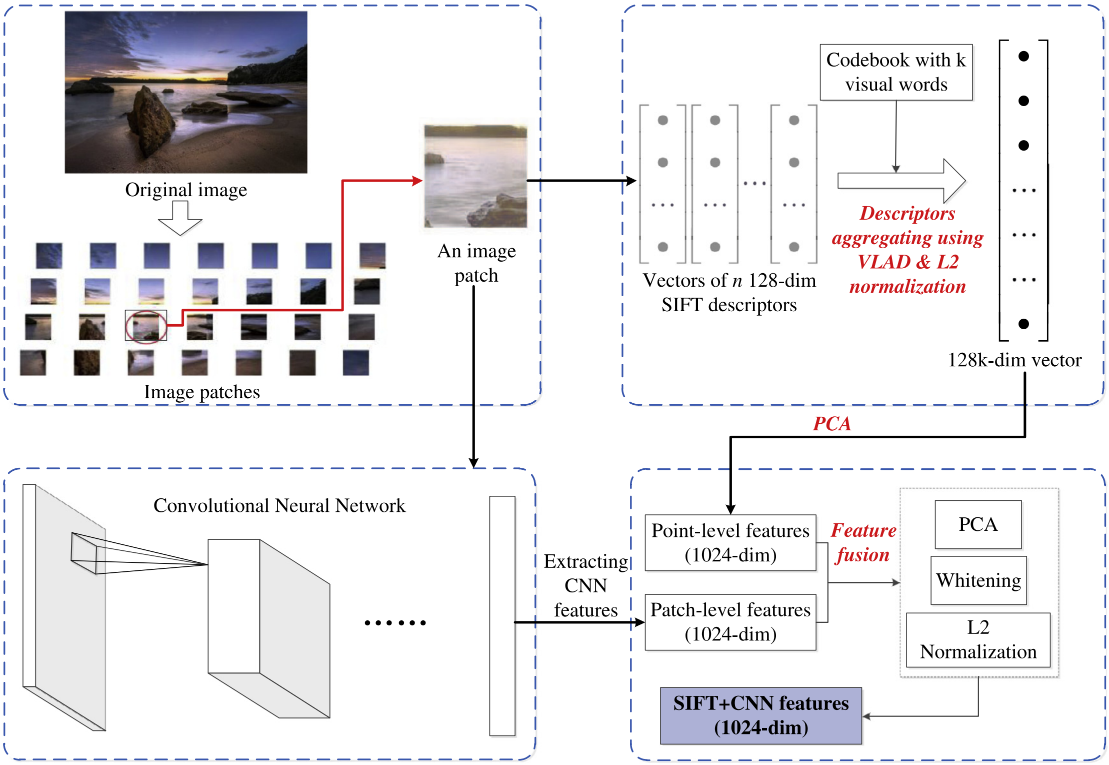

Original image
Image patches
An image patch
Codebook with k visual words
Vectors of n 128-dim SIFT descriptors
Descriptors aggregating using VLAD & L2 normalization
128k-dim vector
PCA
Convolutional Neural Network
Extracting CNN features
Point-level features (1024-dim)
Patch-level features (1024-dim)
Feature fusion
PCA
Whitening
L2 Normalization
SIFT+CNN features (1024-dim)

Fig. 3. Flowchart of SIFT+CNN features generation.

In order to fuse f_v and f_p, we directly concatenate to construct an integrated feature as follows.

$$f = [f_v, f_p] \tag{11}$$

As the integrated feature f (2048-dim) has high dimensionality, PCA and whitening are operated on it as follows (Yan et al., 2016).

$$f_{wht} = diag(1./sqrt(sv_1, sv_2, \cdots, sv_h)) \times M_{PCA} \times \frac{f}{\|f\|_2} \tag{12}$$

where M_{PCA} denotes the matrix of PCA transformation, h refers to the number of reserved dimensions after dimensionality reduction, and sv_i means the ith singular value. Next, L2 normalization is used to achieve a compact vector $f_{final} = \frac{f_{wht}}{\|f_{wht}\|_2}$ with 1024 dimensions.

Based on the above analysis, flowchart of SIFT+CNN features generation is shown in Fig. 3.

It can be observed from Fig. 3 that the original image is firstly divided to several patches. Secondly, for each image patch, its SIFT descriptors are extracted, and then they are aggregated together to a 128k-dim vector, which is also named as the Point-level feature. Thirdly, CNN features (also named as Patch-level feature) are extracted from this image patch. Fourthly, both Point and Patch level features are fused to construct SIFT+CNN features (1024-dim) by some operations, such as PCA, Whitening, and L2 Normalization. As the SIFT+CNN feature contains rich information, we utilize K-means to learn a visual dictionary with 1000 visual words, and then each image patch is represented as a 1000-dim feature (BoVW+SIFT+CNN).

4.3. Formulation of the proposed MMDF-LDA topic model

GPS information are available in Flickr images, for example, the Flickr image has the GPS latitude and longitude in its EXIF (shown in Fig. 1). Therefore, we use the geographical topic, which has been discussed by Yin, Cao, Han, Zhai, & Huang (2011), to combine geographical clustering and topic model by modeling the geographical distribution of different types of metadata. The proposed model follows the intuition that metadata of a social image rely on its location and topic, while topics have different distributions over different geographical regions.

Definition 4 (**Geographical topic**). A geographical topic refers to a spatially coherent meaningful theme, and the words with high concurrence degree in space are possible to be clustered in a topic. A word has different probability to be exploited in different geographical regions, and different regions have different distributions of visual patterns. Furthermore, each region has its own geographical topic distribution.

The proposed method is made up of three components: (1) the MMDF-LDA model, (2) Training process, and (3) Testing process. The key component of our method is the MMDF-LDA model, which is trained by inferring latent variables conditioned on observed variables using the collapsed Gibbs sampling. Particularly, four different data modalities are extracted from social images, such as Visual words, Tags, Comments, and Geographic information. The test social image is divided into several patches, and patch level annotations are obtained by estimating the probability of annotation-patch assignment. An annotation with the highest assignment probability is assigned to an image patch. Afterwards, all annotations are ranked according to the frequency of annotation-patch assignment, and top ranked ones are reserved as the final annotations. Framework of the proposed model is shown in Fig. 4.

It Fig. 4, parameters R, N, K, A, H, T, C, G denote the number of latent regions, social images, topics, annotations, image patches, tags, comment words, and geo tags in the training dataset respectively, and other notations in this figure are illustrated in Table 1 as follows.

Suppose that the annotation set of social image I_i is represented as $A_i = \{a_{i1}, a_{i2}, \cdots, a_{i|A_i|}\}$. Similarly, the image patch set, the tag set, the comment word set and the geo tag set of I_i are denoted as $H_i = \{h_{i1}, h_{i2}, \cdots, h_{i|H_i|}\}$, $T_i = \{t_{i1}, t_{i2}, \cdots, t_{i|T_i|}\}$, $C_i = \{c_{i1}, c_{i2}, \cdots, c_{i|C_i|}\}$, and $G_i = \{g_{i1}, g_{i2}, \cdots, g_{i|G_i|}\}$ respectively. Loca-

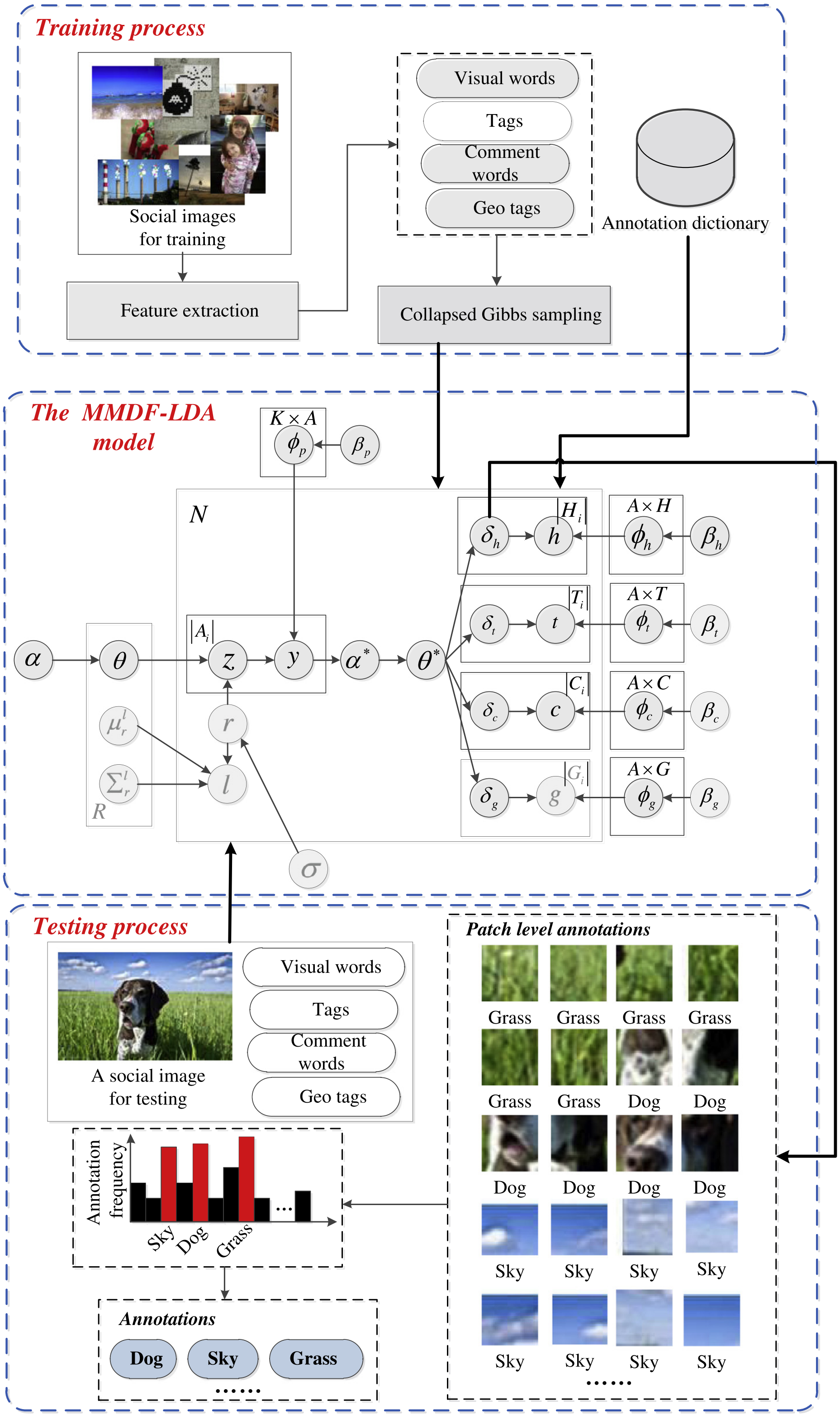


Fig. 4. Framework of the proposed social image annotation method.

tion of I_i is represented as a two-dimensional vector $l_i=\{l_{la}^i, l_{lo}^i\}$, where l_{la}^i and l_{lo}^i denote the latitude and longitude respectively, and we use K-means to cluster all training images into R latent regions according to their locations. Each social image belongs to a region, and each geographical region has its own topic distribution.

Based on the above definitions, the generative process of our proposed MMDF-LDA model is given as follows.

The generative process of the MMDF-LDA model
For each social image I_i in the training dataset

(1) **Geographical region part**

1) **For** each geographical region r_j in social image I_i
Sampling a discrete region distribution $r_j \sim Discrete(\sigma)$

2) Sampling a location l_i from the normal distribution $N(u_{rj}^l, \sum_{rj}^l)$.

Table 1
Notation descriptions.

Notation	Descriptions	Notation	Descriptions
α	Hyper-parameter for θ	β_t	Hyper-parameter for ϕ_t
θ	Region-specific topic distribution	ϕ_t	Annotation distribution over tags
z	Variables that allocate topics to annotations	t	Tag of a social image
y	Annotation of the social image	δ_t	Variables that allocate labels to tags
α^*	Hyper-parameter for θ^*	β_c	Hyper-parameter for ϕ_c
θ^*	Annotation distribution for a social image	ϕ_c	Annotation distribution over comments
β_p	Hyper-parameters for ϕ_p	c	Comment word of a social image
ϕ_p	Topic distribution over annotations	δ_c	Variables that allocate labels to comment words
β_v	Hyper-parameter for ϕ_v	β_g	Hyper-parameter for ϕ_g
ϕ_h	Annotation distribution over image patches	ϕ_g	Annotation distribution over geo tags
h	Image patch of a social image	g	Geo tag of a social image
δ_h	Variables that allocate labels to image patches	δ_g	Variables that allocate labels to geo tags
μ_r^l	Mean vector of region r	$\sum_r^l$	Covariance matrix of region r
l	Geographical location of a social image	σ	Geographical region importance

```
        3) End for
    (2) Sampling a topic distribution of annotations $\theta \sim Dir(\alpha)$
    (3) Annotation part
        1) For each annotation $a_j$ of $A_i$ in social image $I_i$
        2) Sampling a topic allocation $z_j = Mul(\theta)$
        3) Sampling an annotation from $p(a|z_j, \phi_p)$=$Mul(\phi_p^{z_i})$ from the $j^{th}$ topic.
        4) End for
    (4) Sample a distribution for an annotation $\theta^* \sim Dir(\cdot|\alpha_i^*)$
    (5) Image patch part
        1) For each image patch $h_j$ of the social image $I_i$
        2) Sampling an annotation allocation $\delta_{hj} \sim Mul(\theta^*)$
        3) Sampling an image patch by $p(h|\delta_{hj}, \phi_h) = \prod_{h=1}^{D} Mul((\phi_h^{\delta_{hj}}))$, $\phi_h^{\delta_{hj}}$ refers to a $D$ dimensional multinomial for annotation $\delta_{hj}$
        4) End for
    (6) Tag part
        1) For each tag $t_j$ of the social image $I_i$
        2) Sampling an annotation allocation $\delta_{tj} \sim Mul(\theta^*)$
        3) Sampling a tag by $p(t|\delta_{tj}, \phi_t) = Mul((\phi_t^{\delta_{tj}}))$
        4) End for
    (7) Comment word part
        1) For each comment word $c_j$ of the social image $I_i$
        2) Sampling an annotation allocation $\delta_{cj} \sim Mul(\theta^*)$
        3) Sampling a comment word by $p(c|\delta_{cl}, \phi_c) = Mul((\phi_c^{\delta_{cl}}))$
        4) End for
    (8) Geo tag part
        1) For each geo tag $g_j$ of the social image $I_i$
        2) Sampling an annotation allocation $\delta_{gj} \sim Mul(\theta^*)$
        3) Sampling a geo tag by $p(w|\delta_{gj}, \phi_g) = Mul((\phi_g^{\delta_{gj}}))$
        4) End for
End for
```

4.4. Model inference

The proposed method highly relies on the model inference, and several methods have been proposed to estimate the latent variables in a probabilistic graphical model. However, exact inference is intractable in the topic model. Hence, appropriate schemes should be utilized, such as Variational inference (Blei et al., 2003) and Gibbs sampling (Griffiths & Steyvers, 2004). Gibbs sampling yields relatively simple algorithm for high dimensional data modeling, and it also exploits the Markov chain to simulate the generation processes of the topic model. In this section, we discuss how to exploit the collapsed Gibbs sampling to estimate latent variables conditioned on the observed variables.

There are four modalities in the proposed MMDF-LDA model, that is, (1) Modality 1: Visual feature (M_1), (2) Modality 2: Tag (M_2), (3) Modality 3: User comment (M_3), and (4) Modality 4: Geo tag (M_4).

For image I_i, a latent region r_j can be drawn from the distribution as follows.

$$P(r_j|\sigma)P\left(l_i \middle| u_{rj}^l, \sum_{rj}^l\right)P(h_i|r_j)P(t_i|r_j)P(c_i|r_j)P(g_i|r_j) \tag{13}$$

where $P(l_i|u_{rj}^l, \sum_{rj}^l)$ denotes the probability density function for a multivariate normal distribution which is related to region r_j. Furthermore, $P(r_j|\sigma)$, u_{rj}^l and $\sum_{rj}^l$ are estimated as follows.

$$P(r_j|\sigma) = \frac{\sum_{I_i \in I} d(I_i \to r_j) + \sigma}{|I| + \sigma|R|} \tag{14}$$

$$u_{rj}^l = \frac{\sum_{I_i \in I} d(I_i \to r_j) l_i}{N_{I_i}^{r_j}} \tag{15}$$

$$\sum_{rj}^l = \frac{\sum_{I_i \in I} d(I_i \to r_j)(l_i - u_{rj}^l)^T (l_i - u_{rj}^l)}{N_{I_i}^{r_j} - 1} \tag{16}$$

where $d(I_i \to r_j)$ is set to 1 only if image I_i is assigned to latent region r_j, and $N_{I_i}^{r_j}$ refers to the number of images assigned to r_j. In addition, $P(h_i|r_j)$ is estimated as follows.

$$P(h_i|r_j) = \prod\nolimits_{h_{im}} P(h_{im}|r_j) \tag{17}$$

where h_{im} is the mth dimension in the feature vector of h_i. Additionally, $P(t_i|r_j)$, $P(c_i|r_j)$, and $P(g_i|r_j)$ can be estimated just like $P(h_i|r_j)$.

For the image patch modality, collapsed Gibbs sampling is used to update annotation assignment for the jth patch (h_j) of image I_i as follows.

$$P\left(a \to h_j \middle| h_j, I_{\neg j}, \delta_{h\neg j}, t, \delta_t, c, \delta_c, g, \delta_g, \alpha_i^*, \beta_h\right) \propto \frac{N_{a,i,\neg j}^{M_1} + \sum_{q=2}^4 N_{a,i}^{M_q} + \alpha_{i,a}^*}{N_{i,\neg j}^{M_1} + \sum_{q=2}^4 N_i^{M_q} + \sum_b \alpha_{i,b}^*} \times \frac{\prod_{v=1}^V \prod_{y=1}^{N_{jiv}} \left(N_{v,a,\neg j}^{M_1} + N_{jiv} + \beta_h\right)}{\prod_{z=1}^{N_{ji}} \left(N_{a,\neg j}^{M_1} + N_{ji} + V \cdot \beta_h\right)} \tag{18}$$

where $a \to h_j$ denotes that assigning annotation a to image patch h_j, $I_{\neg j}$ refers to images in training dataset by excluding images which contain image patch h_j, $\delta_{v\neg j}$ is the variable which assigns annotations to image patches excluding h_j, $N_{a,i,\neg j}^{M_q}$ refers to the number of times that annotation a is assigned to the qth modality of I_i excluding h_j, and $N_{i,\neg j}^{M_q}$ is the number of times that all annotations are assigned to the qth modality of I_i excluding h_j. Additionally, N_{ji} is the number of visual words in image patch h_j of I_i, N_{jiv} denotes the number of visual words v in image patch h_j of I_i, and V refers to the total number of visual words.

Eq. (18) demonstrates that the annotation assignment of image patches is determined by integrating all types of data modality of social images, that is to say, the proposed method annotates social images by multi-modal data fusion. Just like updating the annotation assignment for image patches (shown in Eq. (18)), we also can obtain annotation assignment for other modalities.

After finishing the collapsed Gibbs sampling, the posterior distribution of annotations (θ_i^*) for social image I_i can be estimated as follows.

$$\theta_{ia}^{*}=\frac{\sum_{q=1}^{4} N_{a,i}^{M_q}+\alpha_i^{*}}{\sum_{q=1}^{4} N_i^{M_q}+\sum_b \alpha_{i,b}^{*}} \tag{19}$$

As the above four modalities play different roles in social image annotation, we should weight some modalities more than others. To achieve this goal, we should set different weights for assignment counting of different modalities. Then, $\sum_{q=1}^{4} N^{M_q}$ in Eqs. (18) and (19) are replaced by the following equation.

$$\sum_{q=1}^{4} L_q \cdot \sum_{q=1}^{4}\left(\lambda_q \cdot \frac{N^{M_q}}{L_q}\right) \tag{20}$$

$$\text{s.t.} \quad \sum_{q=1}^{4}\lambda_q=1, \quad \text{and} \quad \lambda_q \in [0,1]$$

where L_q is the length of the q^{th} modality, and λ_q is the weighting coefficient.

Afterwards, parameter ϕ_h are estimated as follows.

$$\widehat{\phi}_h^{kj} = \frac{N_k^j+\beta_h}{\sum_{j=1}^{H}\left(N_k^j+\beta_h\right)} \tag{21}$$

where N_k^j refers to the times of image patch h_j assigned to the kth topic. Furthermore, ϕ_t,ϕ_c,ϕ_g,ϕ_p are estimated similarly to ϕ_h.

The time complexity of the model inference process is as follows.

$$O(N_{iter}RN_T(N_{pt}AVV_t + TA + CA + GA + KA)) \tag{22}$$

where N_{iter} is the number of iterations, R is the number of geographical regions, N_T is the number of social images in the training dataset, N_{pt} is the number of patches of a social image, V is the number of visual words which appear at least once in an image patch, V_t indicates the number of occurrences of a specific visual word in an image patch. In addition, A, T, C, G, and K denote the number of annotations, tags, comment words, geo tags, and topics respectively.

Particularly, O($N_{pt}AVV_t$), O(TA), O(CA), O(GA) and O(KA) denote the time complexity of sampling annotations for image patches, tags, comment words, geo tags, and annotations respectively.

5. Experiment

To demonstrate the effectiveness of our proposed model, we compare it with other methods on both non-social image datasets and social image datasets. Social images used in this paper are collected from Flickr, which is a famous Web 2.0 photo sharing website.

5.1. Dataset

To make the experimental results more persuasive, four different types of image datasets are chosen, that is, (1) Corel5K (Duygulu, Barnard, de Freitas, & Forsyth, 2002), (2) MIRFlickr-25K (Huiskes & Lew, 2008), (3) NUS-WIDE (Chua et al., 2009), and (4) Flickr-MMM (newly constructed by us).

The publicly available Corel5K dataset is a non-social image dataset with 5000 images (4500 training images and 500 testing images). Corel5K is made up of 50 image classes, with 100 images per class. Each image is divided into a set of 20 × 20 patches via a sliding window with the twenty pixels interval. As the resolutions of images in Corel5K are 192 × 128 or 128 × 192, approximately 54 20 × 20 patches can be obtained. Moreover, each image in Corel5K is annotated with 1 to 5 labels, and the average number of labels for one image is nearly 3.22. We choose 260 labels in our experiment which have been annotated on both training and testing images. As the Corel5K dataset does not have any other metadata except for visual features, in this experiment, we only utilize visual features to annotate images in Corel5K.

MIRFlickr-25K is used as a standard social image dataset, and it includes 25,000 social images which are collected from Flickr with 1386 tags. For each social image in MIRFlickr-25K, tags and EXIF information are available. We limit the set of tags to the 457 most frequent ones which appear at least 50 times in MIRFlickr-25K.

NUS-WIDE denotes a Web image dataset created by NUS's Lab for Media Search. This dataset includes 269,648 images and the associated tags from Flickr, with a total of 5,018 unique tags. Moreover, there are 81 concepts in NUS-WIDE. Particularly, as 500-dim bag of visual words based on SIFT descriptions are provide by NUS-WIDE, in this experiment, we directly use these 500-dim bags of visual words as the BoVW+SIFT features.

However, MIRFlickr-25K and NUS-WIDE do not contain multi-modal metadata of social images. To test the annotation performance in the multi-modal metadata environment, we collect 5000 Flickr images with rich metadata through the Flickr public API to construct a new dataset - Flickr images with multi-modal metadata (denoted as Flickr-MMM). In Flickr-MMM, metadata of social images contain: (1) User-supplied tags, (2) User comments, (3) Geographic tags, and (4) Coordinate of the location where the photo is taken. Words in User-supplied tags, User comments and Geographic tags are processed by removing stop words and correcting spelling errors. Moreover, we check all words with Wikipedia, and words that do not exist in Wikipedia are deleted. After the above process, we construct an annotation dictionary for Flickr-MMM with 367 words.

Specifically, ten image categories are contained in Flickr-MMM, including: (1) Flower, (2) Dog, (3) Beach, (4) Bird, (5) Car, (6) Street, (7) Lake, (8) Eiffel tower, (9) Pyramid, and (10) Great wall. Each image category contains 500 images, among which we randomly choose 50 images to be testing images. Similar to the Corel5K dataset, 4500 images in Flickr-MMM are utilized for training and others are exploited for testing.

As images in different datasets have different resolutions, for the sake of fairness, we set the number of patches to approximately 400 for the images in MIRFlickr-25K, NUS-WIDE and Flickr-MMM. Hence, suppose that the height and the width of an image are ht and wd respectively, the size of each patch is set as $\lfloor\sqrt{(ht \times wd)/400}\rfloor \times \lfloor\sqrt{(ht \times wd)/400}\rfloor$.

5.2. Performance evaluation criteria

Average precision ratio (APT) and Average recall ratio (ART) (Zhang et al., 2013) are used to testify the performance of semantic information mining, which are defined as follows.

$$APT = \frac{1}{W} \cdot \sum_{i=1}^{W} \frac{|CA(w_i)|}{|TA(w_i)|} \tag{23}$$

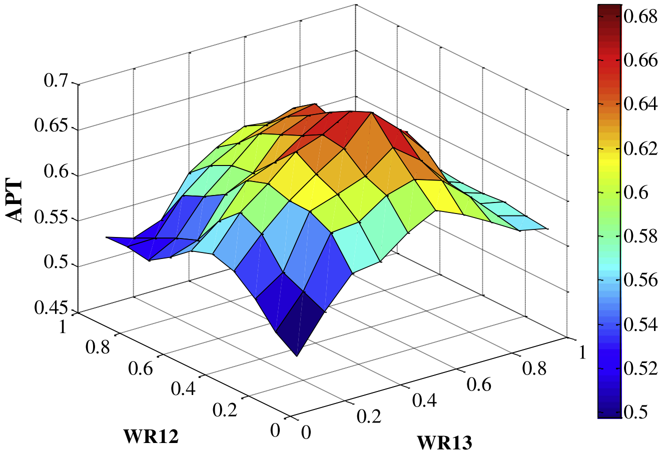


Fig. 5. WR_{12} vs. WR_{13}.

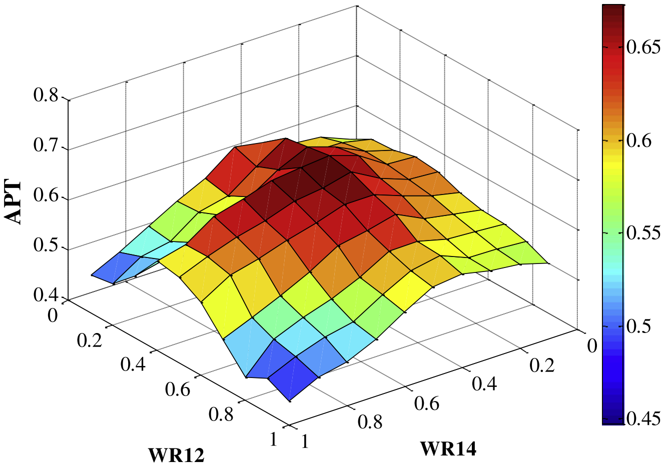


Fig. 6. WR_{12} vs. WR_{14}.

$$ART = \frac{1}{W} \cdot \sum_{i=1}^{W} \frac{|CA(w_i)|}{|MA(w_i)|} \tag{24}$$

where $CA(w_i)$ refers to the set of social images which are correctly annotated by the word w_i with the annotating algorithm, $TA(w_i)$ is the set of social images which have been annotated by w_i with the annotating algorithm, and $MA(w_i)$ denotes the set of social images which are manually annotated by w_i. Specifically, W is the total number of words in the annotation dictionary.

Afterwards, to test the quality of the social image annotations, Top N precision ($P@N$) and Top N coverage ($C@$N) are utilized as well. $P@N$ is used to measure the precision of the top N ranked annotations for an image, which is defined as follows.

$$P@N = \frac{1}{|I^T|} \cdot \sum_{I_i \in I^T} \left(\frac{1}{N} \cdot |PT_N(I_i)| \right) \tag{25}$$

where I^T denotes the testing social image dataset, and $PT_N(I_i)$ refers to the set of the words, which are correctly assigned to social image I_i.

Similarly, $C@N$ denotes the ratio of social images which are correctly annotated by at least one annotation in the top N ranked annotations.

$$C@N = \frac{1}{|I^T|} \cdot \sum_{I_i \in I^T} T_{I_i}(N) \tag{26}$$

where $T_{I_i}(N)$ is set to 1 if there is at least one correct annotation in the top ranked annotations of I_i, otherwise, $T_{I_i}(N)$ is set to zero.

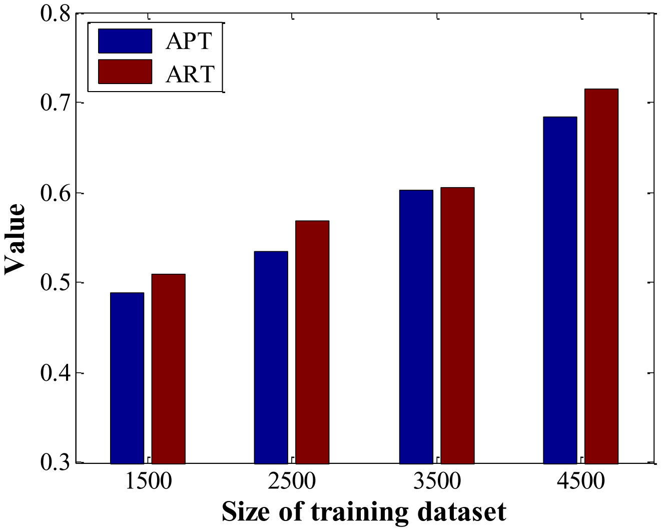


Fig. 7. Influences of different size of the training dataset.

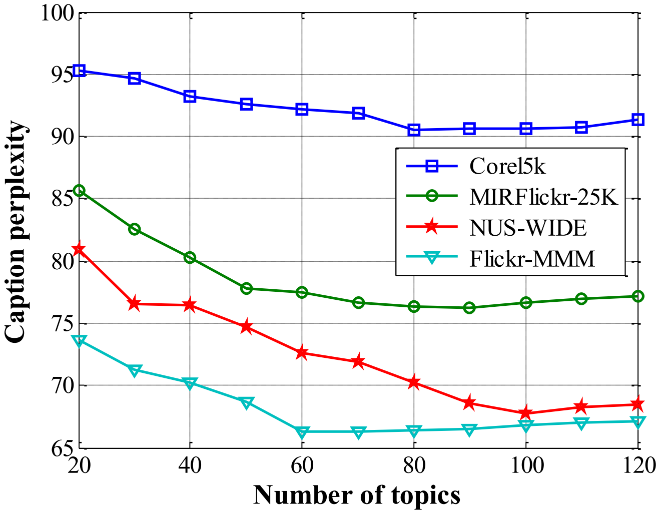


Fig. 8. Caption perplexity with different number of topics.

F1 value combines the precision and recall together, and it is defined as follows.

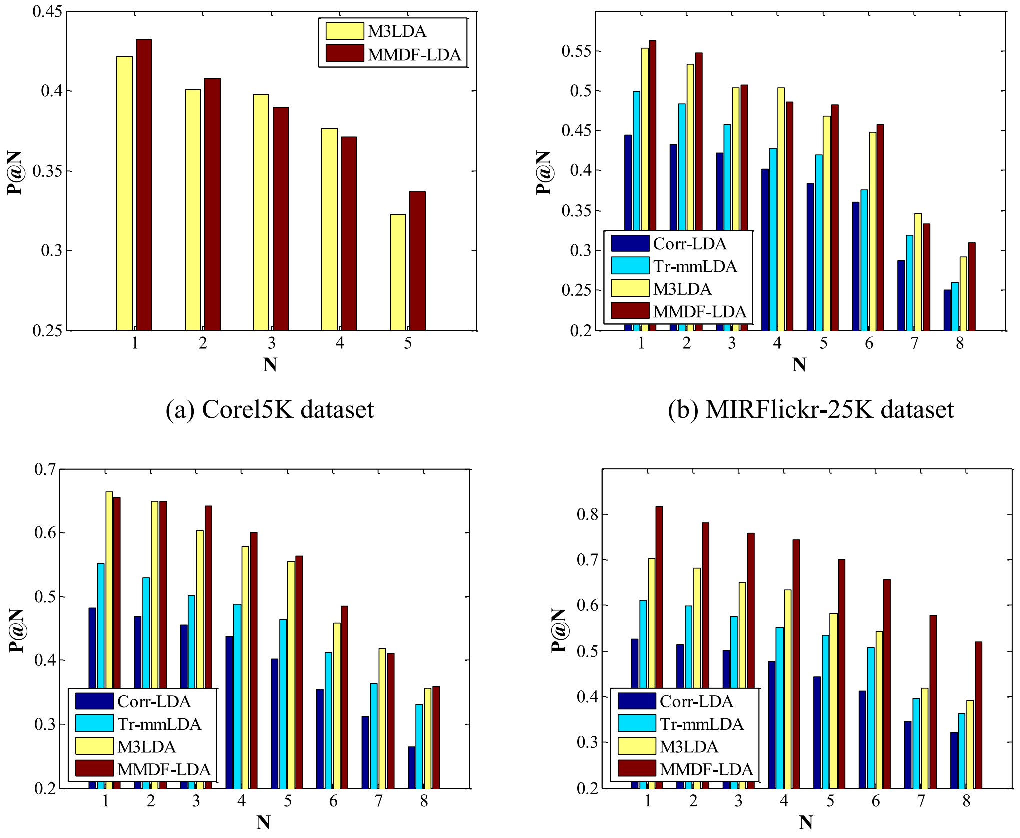
$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{27}$$

The higher score of F1 means the better performance of the social image annotation.

Furthermore, we also adopt the caption perplexity (Blei & Jordan, 2003; Putthividhya et al., 2010b) as a performance metric. The essential quantity that should be calculated is the conditional probability of caption words (annotations) given a testing image $p(a_{ij}|I_i)$. The caption perplexity is defined as follows.

$$Perp = \exp\left(-\sum_{i=1}^{|I^T|} \sum_{j=1}^{|A_i|} \log p\big(a_{ij}|I_i\big) / \sum_{i=1}^{|I^T|} |A_i| \right) \tag{28}$$

where A_i is the annotation set of testing image I_i, and a_{ij} refers to the jth annotation of I_i. Perplexity is indeed the inverse of the geometric mean likelihood, which implies that the model which gives higher conditional likelihood will lead to a lower perplexity (that is, lower numbers are better) (Putthividhya, Attias, & Nagarajan, 2010).



(a) Corel5K dataset

(b) MIRFlickr-25K dataset

(c) NUS-WIDE dataset

(d) Flickr-MMM dataset

Fig. 9. Performance evaluation using the P@N metric with BoVW+SIFT+CNN.

Table 2
Data modality settings for different datasets.

Dataset	M_1	M_2	M_3	M_4	Geographical topic
Corel5K	√				
MIRFlickr-25K	√	√			
NUS-WIDE	√	√			
Flickr-MMM	√	√	√	√	√

Table 3
Performance comparison for different visual features.

Dataset	APT			ART		
	vf1	vf2	vf3	vf1	vf2	vf3
Corel5K	0.355	0.36	0.372	0.377	0.387	0.398
MIRFlickr-25K	0.432	0.441	0.454	0.441	0.454	0.466
NUS-WIDE	0.561	0.568	0.591	0.588	0.608	0.624
Flickr-MMM	0.684	0.697	0.719	0.715	0.735	0.757

5.3. Experimental results and analysis

In this experiment, Correspondence LDA (Corr-LDA) (Blei & Jordan, 2003), Topic-regression multi-modal LDA (Tr-mmLDA) (Putthividhya et al., 2010b), and Multimodal Multi-instance Multi-label LDA (M3LDA) (Nguyen et al., 2013) are compared with our proposed model.

Particularly, we set hyperparameters according to empirical values:α=0.001,β_p=0.01, β_t=0.001, β_c=0.001, β_g=0.001. Parameter σ is estimated by computing the proportion of social images in a given geographical region to the whole images. The number of iterations for model training and testing are set to 2000 and 100 respectively. For the Flickr-MMM dataset, the number of latent regions is set to 20.

5.3.1. Parameters selection

In this subsection, BoVW+SIFT features are used to conduct parameters selection.

(1) **Weighted fusion coefficient**

As different modality of metadata may have different influences in the process of semantic annotation, we should try to estimate the weight for each modality. Four modalities used in this paper are illustrated in Section 4.4.

As is shown in Eq. 20, weight of the modality M_i is represented as λ_i. The weighted ratio between λ_i and λ_j are defined as follows.

$$WR_{ij} = \frac{\lambda_i}{\lambda_i + \lambda_j} \tag{29}$$

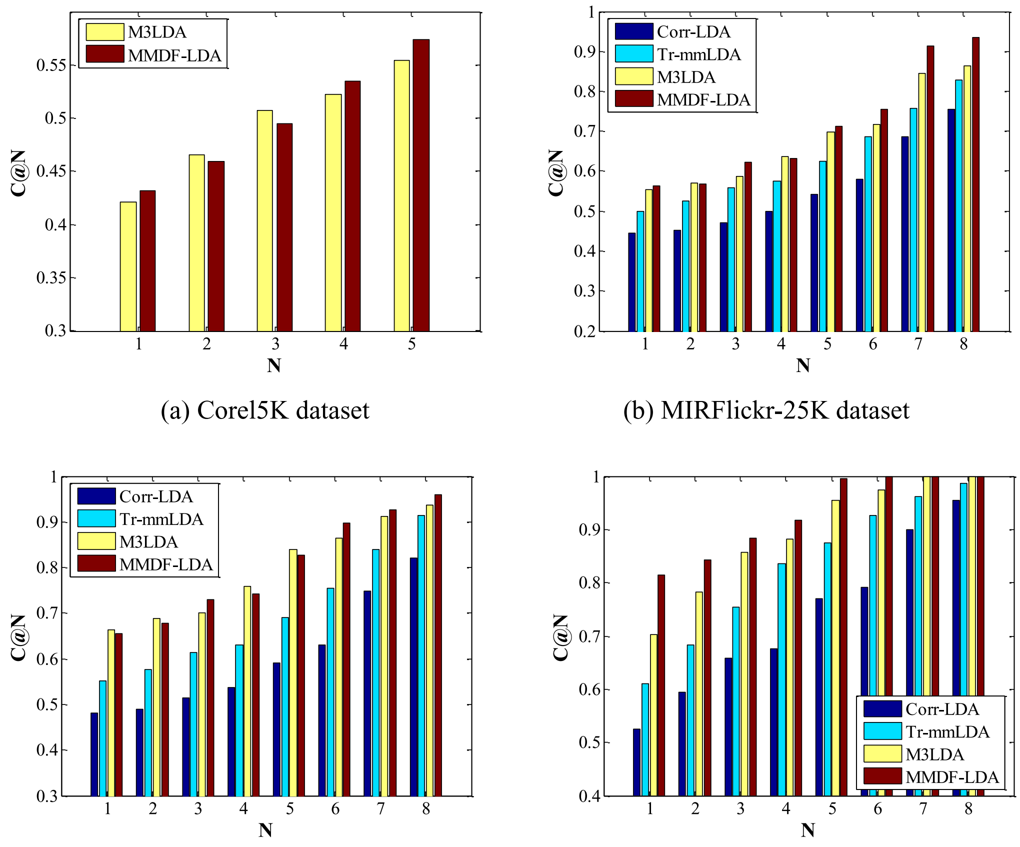
To illustrate the influences of different weighted fusion coefficients for annotating results, we test different pairs of weighed ratios in Figs. 5 and 6 (the number of final annotations are set to 4).

Fig. 5 shows that the highest value of APT is achieved when WR_{12} and WR_{13} are set to 0.4 and 0.6 respectively.

Fig. 6 illustrates that WR_{12} and WR_{14} should be set to 0.4 and 0.4 respectively to obtain the best performance.

Integrating experimental results in Figs. 4 and 5, WR_{12}, WR_{13}, WR_{14} should be set to 0.4, 0.6, and 0.4 respectively. Therefore, weighted fusion coefficients λ_1,λ_2,λ_3,λ_4 are 0.214, 0.318, 0.144 and 0.324 respectively.

(2) **Size of the training dataset**



(a) Corel5K dataset

(b) MIRFlickr-25K dataset

(c) NUS-WIDE dataset

(d) Flickr-MMM dataset

Fig. 10. Performance evaluation using the C@N metric with BoVW+SIFT+CNN.

We test the influence of the size of training dataset on Flickr-MMM by using 500 images for testing and then change the size of training dataset from 1500 to 4500. Our programs run on a computer with Intel Core i7-6700 processor (3.4 GHz) and 8GB Memory. When the scale of training dataset is 4500, each collapsed Gibbs sampling iteration consumes about 38.8 seconds.

Fig. 7 demonstrates that both APT and ART increase when the size of the training dataset increases. In addition, for each testing image, the annotating processing consumes approximately 1.2 seconds.

(3) **Topic number**

To avoid overfitting, we use the caption perplexity (Blei & Jordan, 2003) to test the quality of image annotating results with different number of topics. Perplexity, which is used in the language modeling community, is equivalent algebraicly to the inverse of the geometric mean per-word likelihood (Blei & Jordan, 2003). Lower value of the caption perplexity means better performance.

Fig. 8 shows that overfitting is not serious in our proposed model. To achieve higher quality of annotating results and lower computation cost, the number of topics for Corel5K, MIRFlickr-25K, NUS-WIDE and Flickr-MMM are set to 80, 50, 100 and 60, respectively.

5.3.2. Performance of social image annotation

After the parameters are determined, we test the performance of different methods using the proposed four datasets: (1) Corel5K, (2) MIRFlickr-25K, (3) NUS-WIDE, and (4) Flickr-MMM. However, four datasets we used have different data modality settings (shown in Table 2).

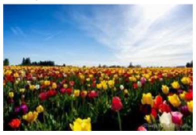
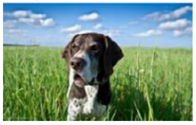
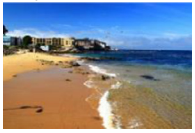
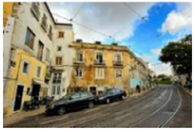
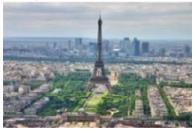
In Table 2, symbol $\surd$ means that a given dataset has the corresponding item in this experiment.

Firstly, we validate the effectiveness of three different visual feature descriptions (mentioned in Section 4.2): (1) BoVW+SIFT (vf1), (2) DBoVW+SIFT (vf2), and (3) BoVW+SIFT+CNN (vf3). To extract the CNN features, we implement the GoogLeNet using the public Caffe (Convolutional Architecture for Fast Feature Embedding) toolbox, which is a deep learning framework, originally developed at UC Berkeley. All image patches should be resized before inputting to GoogLeNet. Furthermore, to improve the generalization ability of the CNN model, we use a pre-trained model (bvlc_googlenet.caffemodel) that has been optimized on the ImageNet. In particular, bvlc_googlenet.caffemodel can be freely downloaded from the Internet. Table 3 lists the APT and ART values on various datasets with various visual features.

From Table 3, we can find that BoVW+SIFT+CNN performs the best, because CNN features can capture high level semantic information, and integrating CNN features with SIFT descriptors may well complement each other for the social image annotation task. In addition, DBoVW+SIFT is superior to BoVW+SIFT, the major reason is that DBoVW can effectively distinguish the most descriptive visual words and remove noise data.

Secondly, to make a fair comparison, we exploit BoVW+SIFT+CNN as the visual feature descriptor to test the performance of all methods. As M3LDA and MMDF-LDA model permit the modality missing, we only utilize these two methods to annotate Corel5K images (with only visual features). For the Flickr-MMM dataset, Corr-LDA, Tr-mmLDA, and M3LDA can only use M_1 and M_2, because these methods are developed to support merely two types of data modalities. Moreover, we test the per-

Table 4
Examples of image annotating results.

Social image	User-supplied tags	Corr-LDA	Tr-mmLDA	M3LDA	MMDF-LDA
Image 1 	Tulips, Flowers, Flower fields, Oregon, Woodburn, Spring	Flower, Sky, Sea, Park, Outdoor, Day, Travel, Spring	Flower, Sky, Field, Park, Cloud, Outdoor, Blue, Food	Plant, Flower, Sky, Cloud, Park, Outdoor, Adult, Bee	Tulip, Flower, Oregon, Woodburn, Sky, Landscape, Park, Outdoor
Image 2 	Pero, Dog	Dog, Sky, Green, Play, Baby, Nature, Park, Landscape	Nature, Dog, Child, Cloud, Sky, Tree, Animal, Color	Dog, Animal, Blue, Happy, Travel, Puppy, Lawn, Water	Dog, Grass, Pet, Animal, Plant, Sky, Cloud, Outdoor
Image 3 	Monterey, Blue, Sky, Sea, Seagull, Rock, 2011, Oct, Beach, Sun, Cloud, Canont2i, California, ...	Sea, Summer, Wave, Boat, Girl, Sun, Tree, Day	Beach, Ocean, People, Sand, Sunset, Walk, Fun, Holiday	Sky, Beach, Boat, Ship, River, Cloud, Nature, Harbor	Sky, Sea, Seagull, Rock, Beach, Cloud, California, Water
Image 4 	Street	Road, Street, Urban, Sky, People, Car, Highway, Traffic	Car, Road, Forest, People, Building, Wall, Travel, Family	Street, Tree, Car, Autumn, Leaf, Road, Sun, Flower	Street, Car, Building, Outdoor, Portugal, Lisbon, Road, Tree
Image 5 	Paris, Skyline, Travel, Tower, Eiffel, Europe, Tourist, Love, Marriage, Engagement, France, Canon, ...	Sky, Building, Tower, Sun, Park, Flower, Bridge, Water	Tower, Landscape, Fall, Tree, Light, Nature, Mountain, Lake	Tower, Eiffel, Sky, Grass, Cloud, Spring, Forest, Street	Tower, City, Eiffel, Paris, France, Landmark, Building, Nature

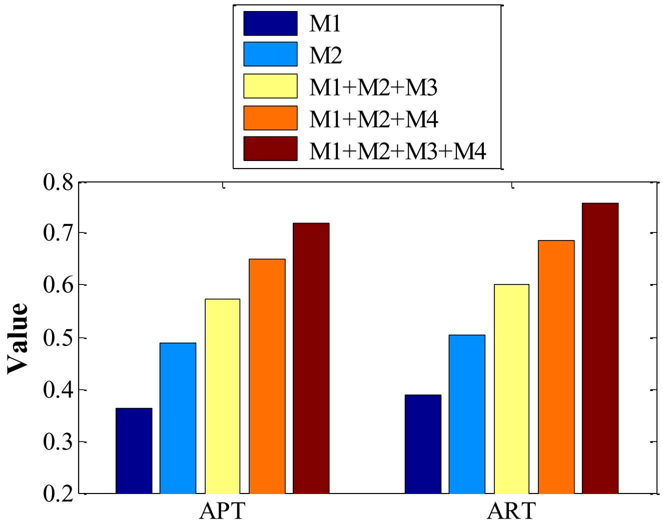


Fig. 11. Performance comparison with different data modalities.

formance of MMDF-LDA without geographical topics when using Corel5K, MIRFlickr-25K, and NUS-WIDE, therefore, we modify the MMDF-LDA model by generating topics from social images instead of geographical regions.

Experimental results on top N precision and top N coverage are shown in Figs. 9 and 10 respectively.

It can be observed from Figs. 9 and 10 that the performance of MMDF-LDA is just similar to M3LDA and only a little better than Corr-LDA and Tr-mmLDA. Meanwhile, MMDF-LDA is obviously superior to other three methods on Flickr-MMM dataset. The reason is that the advantage of MMDF-LDA is to mine semantic information from social images by fusing multi-modal metadata. In addition, fully integrating multiple types of metadata can construct a more effective semantic description space, and then narrow the “semantic gap”.

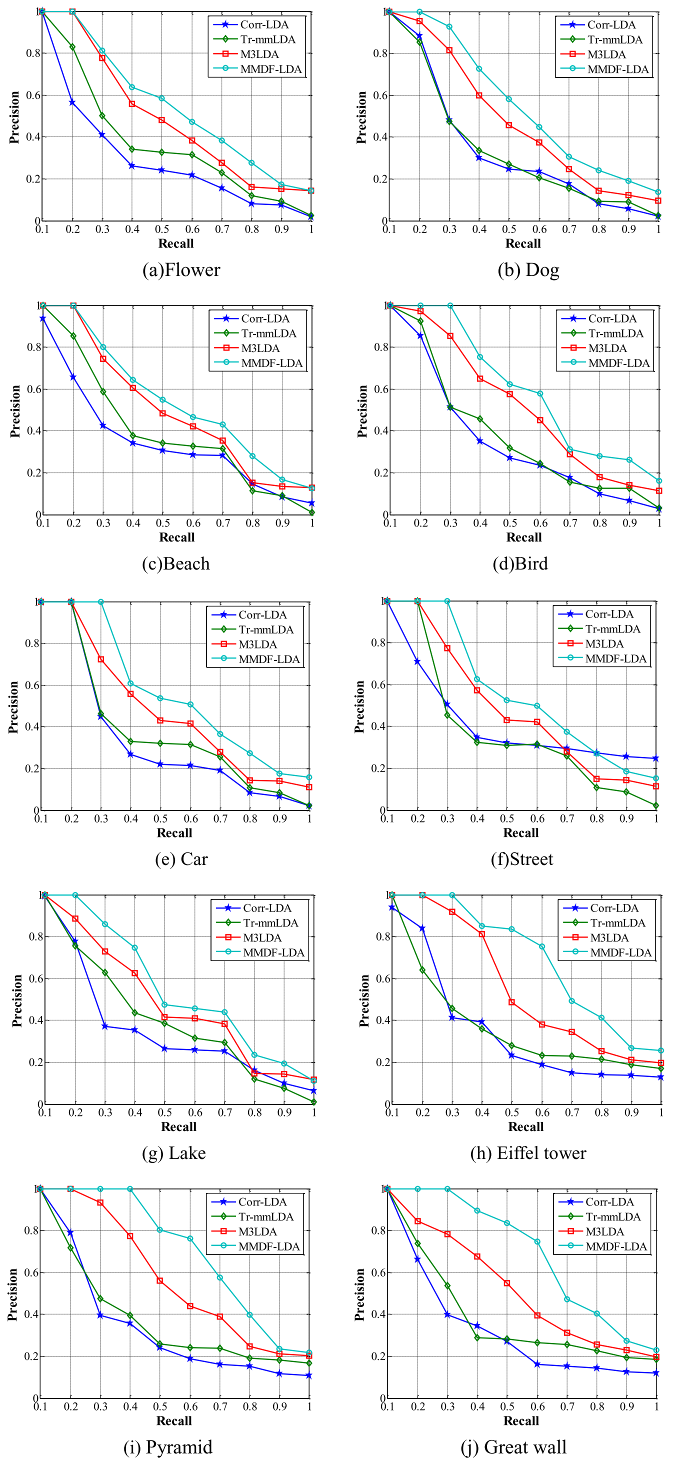
Thirdly, we test the performance of the proposed method on Flickr-MMM with different data modalities using the BoVW+SIFT+CNN feature, and experimental results are listed in Fig. 11.

Fig. 11 shows that when fully integrating all these four types of data modalities ($M_1+M_2+M_3+M_4$), the values of APT and ART can be significantly promoted.

5.3.3. Annotation results

We provide the annotation results of five images from the Flickr-MMM dataset with the top ranked eight annotations (shown in Table 4). Furthermore, user-supplied tags and annotation results obtained from Corr-LDA, Tr-mmLDA and M3LDA are given to make performance comparison with MMDF-LDA.

Table 4 demonstrates that the user-supplied tags are of low quality, e.g. image 2 and image 4 only have two and one tags. In addition, image 3 and image 5 have many noisy words, such as Canont2i, 2011, Oct, and Love. All these methods can provide some meaningful annotations. Especially, more superior quality of annotations are inferred by MMDF-LDA, because MMDF-LDA can not only prune noisy annotations (e.g., Image 3 and Image 5), but also enrich semantics of user-supplied tags (e.g., Image 2 and Image 4) by fusing multi-modal metadata. Furthermore, several words with geographic information are obtained by MMDF-LDA (e.g. Oregon, Woodburn, California, Portugal, Lisbon, Paris and France), which demonstrates that MMDF-LDA can effective find geographical characteristics of social images with location information using the geographical topics. We also observe that social images in different geographical regions have different semantic information and visual patterns.



(a)Flower

(b) Dog

(c)Beach

(d)Bird

(e) Car

(f)Street

(g) Lake

(h) Eiffel tower

(i) Pyramid

(j) Great wall

Fig. 12. Precision-recall curves for ten query words.

5.3.4. Application of the propose method in image retrieval

As image annotation has great significance for text-based image retrieval, we build a text-based social image retrieval system using Flickr-MMM to test the effectiveness of the proposed method with the BoVW+SIFT+CNN feature. Given a single word query, we perform image retrieval by ranking the testing images according to the relevance degree between the image and the query word. The relevance degree between social image I_i and the query word q is defined as follows.

$$rd(I_i, q) = p(q|I_i) \tag{30}$$

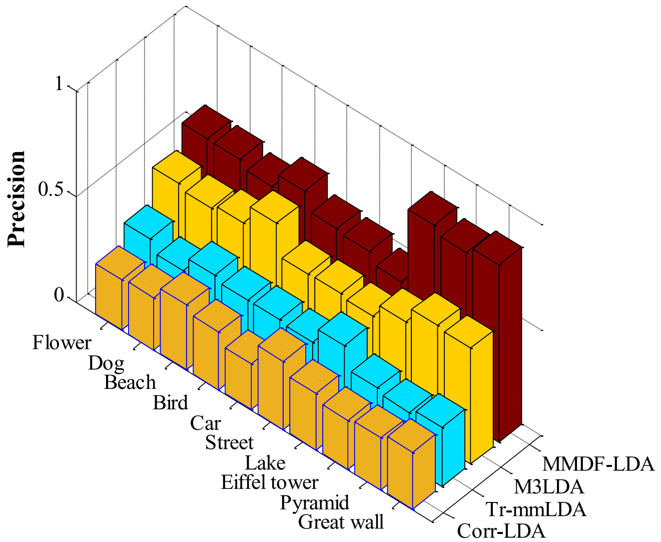


Fig. 13. Precision for all queries (Recall = 0.5).

where $p(q|I_i)$ refers to the conditional distribution, which can retrieve images that are similar to the query word.

Suppose that the annotation results of I_i are represented as $\{a_1, a_2, \cdots, a_n\}$, where the higher ranking annotations have higher relevance to I_i. Then, $p(q|I_i)$ is estimated as follows.

$$p(q|I_i) = \frac{1}{n} \cdot \sum_{i=1}^{n} dis(q, a_i) \tag{31}$$

where the function $dis()$ refers to the Google Similarity Distance (Cilibrasi & Vitanyi, 2007), which calculates the distance between the query word q and the annotation a_i. Precision-recall curves for ten query words in the Flickr-MMM dataset are illustrated in Fig. 12.

It can be observed from Fig. 12 that our proposed method generally yields higher precision at the same recall values for all these ten queries and also achieves a better overall image retrieval performance. Particularly, we also find that the retrieval results for the queries “Eiffel tower”, “Pyramid” and “Great wall” of MMDF-LDA are more accurate than other queries, because our proposed model with the geographical topics can effectively annotate geographical words (e.g. landmark) to images.

Next, we illustrate the precision for all queries when recall is 0.5 (shown in Fig. 13).

Fig. 13 shows that MMDF-LDA outperforms other methods, especially for the landmarks (e.g. Eiffel tower, Pyramid and Great wall).

6. Conclusions

In this work, we aim to annotate social images by fusing multimodal metadata. Our key idea is to learn a probability model to annotate social images by fusing multi-modal metadata together with geographical distribution. In order to achieve this goal, we propose the MMDF-LDA model to mine the relationship between input patterns and output semantics by mapping a high dimensional vector space to a low dimensional vector space. In MMDF-LDA, visual content, user-supplied tags, user comments, and geographic information are fused together. In particular, we suppose that the topics determined by different data modality should be consistent. We conduct extensive experiments using various datasets to validate the effectiveness of the proposed model. We believe that the MMDF-LDA model can provide high quality social annotation results, and also can significantly improve the performance of the text-based image retrieval system.

Next, we list several possible future research directions of this work:

(1) How to guarantee the annotation results with diverse semantics. That is, we should let the top ranked annotations not only be highly relevant to visual content of the image, but also have great semantic compensations with each other.
(2) This work does not consider users' personal preferences in social image annotating. Different users may have different tagging habits and favorite tags. Hence, in the future, we will append user preferences in our model by analyzing users' profiles and interactions between various users.
(3) We will try to partition social images to patches with different scales, and discuss how to select the optimal scale of image patch in the social image annotation task.

Acknowledgements

This work is supported by National Natural Science Foundation of China (61772309, 61303090, 61472221, 61332015), NSFC Joint Fund with Zhejiang Integration of Informatization and Industrialization under Key Project (U1609218), Humanities and Social Sciences Project of Education Ministry (13YJC860023), Science and Technology Project of Ji'nan (201303012) and Youth Science and Technology Star Program of Ji'nan (201406001), Program for Scientific Research Innovation Team in Colleges and Universities of Shandong Province, and Shandong Co-Innovation Center of Future Intelligent Computing. We also appreciate the anonymous reviewers for thoroughly reading the paper and providing thoughtful comments.

References

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE-ACM Transactions on Audio Speech and Language Processing, 22*(10), 1533–1545.

Andrzejewski, D., & Zhu, X. (2009). Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 workshop on semi-supervised learning for natural language processing* (pp. 43–48).

Andrzejewski, D., Zhu, X., & Craven, M. (2009). Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th annual international conference on machine learning* (pp. 25–32). ACM. 2009.

Atrey, P. K., Hossain, M. A., Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems, 16*(6), 345–379.

Bahrololoum, A., & Nezamabadi-pour, H. (2017). A multi-expert based framework for automatic image annotation. *Pattern Recognition, 61*, 169–184.

Berger, C., Voltersen, M., Eckardt, R., Eberle, J., Heyer, T., Salepci, N., & Pacifici, F. (2013). Multi-modal and multi-temporal data fusion: outcome of the 2012 GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6*(3), 1324–1340.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(4–5), 993–1022.

Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 127–134).

Cao, X., Zhang, H., Guo, X., Liu, S., & Meng, D. (2015). Sled: Semantic label embedding dictionary representation for multilabel image annotation. *IEEE Transactions on Image Processing, 24*(9), 2746–2759.

Carson, C., Belongie, S., Greenspan, H., & Malik, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(8), 1026–1038.

Chavali, P., & Nehorai, A. (2014). Hierarchical particle filtering for multi-modal data fusion with application to multiple-target tracking. *Signal Processing, 97*, 207–220.

Chen, X., Flynn, P. J., & Bowyer, K. W. (2006). Fusion of infrared and range data: Multi-modal face images. *Lecture Notes in Computer Science, 3832*, 55–63.

Cheng, Z. Y., Shen, J. L., & Miao, H. Y. (2016). The effects of multiple query evidences on social image retrieval. *Multimedia Systems, 22*(4), 509–523.

Chong, W., Blei, D., & Li, F. F. (2009). Simultaneous image classification and annotation. In *Proceedings of the 2009 IEEE conference on computer vision and pattern recognition* (pp. 1903–1910).

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). NUS-WIDE: A real–world web image database from national university of Singapore. In *Proceedings of the ACM international conference on image and video retrieval* (p. 48).

Cilibrasi, R. L., & Vitanyi, P. M. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 370–383.

Coppock, S., & Mazlack, L. J. (2004). Multi-modal data fusion: A description. *Knowledge-Based Intelligent Information and Engineering Systems, 3214*, 1136–1142.

De, S., Gupta, K., Stanley, R. J., Ghasr, M. T., Zoughi, R., Doering, K., & Palmer, D. D. (2013). A Comprehensive multi-modal NDE data fusion approach for failure assessment in aircraft lap-joint mimics. *IEEE Transactions on Instrumentation and Measurement, 62*(4), 814–827.

Ding, X., Li, B., Xiong, W., Guo, W., Hu, W., & Wang, B. (2016). Multi-instance multi-label learning combining hierarchical context and its application to image annotation. *IEEE Transactions on Multimedia, 18*(8), 1616–1627.

Duygulu, P., Barnard, K., de Freitas, J. F., & Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European conference on computer vision* (pp. 97–112).

Feng, L., & Bhanu, B. (2016). Semantic concept co-occurrence patterns for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(4), 785–799.

Feng, Y., & Lapata, M. (2010). Topic models for image annotation and text illustration. In *Proceedings of the 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 831–839).

Flickr. http://www.flickr.com.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the national academy of sciences: 101* (pp. 5228–5235).

Gu, Y., Qian, X., Li, Q., Wang, M., Hong, R., & Tian, Q. (2015). Image annotation by latent community detection and multikernel learning. *IEEE Transactions on Image Processing, 24*(11), 3450–3463.

Gupta, M., Li, R., Yin, Z., & Han, J. (2010). Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter, 12*(1), 58–72.

Hanbury, A. (2008). A survey of methods for image annotation. *Journal of Visual Languages & Computing, 19*(5), 617–627.

He, C., Zhuo, T., Ou, D., Liu, M., & Liao, M. (2014). Nonlinear compressed sensing-based LDA topic model for polarimetric SAR image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7*(3), 972–982.

Hu, M., Yang, Y., Shen, F., Zhang, L., Shen, H. T., & Li, X. (2017). Robust web image annotation via exploring multi-facet and structural knowledge. *IEEE Transactions on Image Processing, 26*(10), 4871–4884.

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning, 95*(3), 423–469.

Huiskes, M. J., & Lew, M. S. (2008). The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on multimedia information retrieval* (pp. 39–43).

Ireton, M., & Xydeas, C. (1991). Classification of shape for content retrieval of images in a multimedia database. In *proceedings of sixth international conference on the digital processing of signals in communications* (pp. 111–116).

Ivasic-Kos, M., Pobar, M., & Ribaric, S. (2016). Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme. *Pattern Recognition, 52*, 287–305.

Jegou, H., Perronnin, F., Douze, M., Sanchez, J., Perez, P., & Schmid, C. (2012). Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(9), 1704–1716.

Ji, S. W., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 221–231.

Jing, X.-Y., Wu, F., Li, Z., Hu, R., & Zhang, D. (2016). Multi-label dictionary learning for image annotation. *IEEE Transactions on Image Processing, 25*(6), 2712–2725.

Jing, Y., & Baluja, S. (2008). Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(11), 1877–1890.

Jiu, M., & Sahbi, H. (2017). Nonlinear deep kernel learning for image annotation. *IEEE Transactions on Image Processing, 26*(4), 1820–1832.

Lei, C. Y., Liu, D., & Li, W. P. (2016). Social diffusion analysis with common-interest model for image annotation. *IEEE Transactions on Multimedia, 18*(4), 687–701.

Li, D., Ding, Y., Sugimoto, C., He, B., Tang, J., Yan, E., & Dong, T. (2011). Modeling topic and community structure in social tagging: The TTR-LDA-community model. *Journal of the Association for Information Science and Technology, 62*(9), 1849–1866.

Li, F. F., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition* (pp. 524–531).

Li, X., Snoek, C. G., & Worring, M. (2008). Learning tag relevance by neighbor voting for social image retrieval. In *Proceedings of the 1st ACM international conference on multimedia information retrieval* (pp. 180–187).

Li, X., Snoek, C. G., & Worring, M. (2009). Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia, 11*(7), 1310–1322.

Li, X. R., Uricchio, T., Ballan, L., Bertini, M., Snoek, C. G. M., & Del Bimbo, A. (2016). Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys, 49*(1) Article 14.

Li, Z. C., & Tang, J. H. (2017). Weakly supervised deep matrix factorization for social image understanding. *IEEE Transactions on Image Processing, 26*(1), 276–288.

Li, Z. C., Liu, J., Tang, J. H., & Lu, H. Q. (2014). Projective matrix factorization with unified embedding for social image tagging. *Computer Vision and Image Understanding, 124*, 71–78.

Lienou, M., Maitre, H., & Datcu, M. (2010). Semantic annotation of satellite images using Latent Dirichlet Allocation. *IEEE Geoscience and Remote Sensing Letters, 7*(1), 28–32.

Liu, B., Yuan, Q., Cong, G., & Xu, D. (2014). Where your photo is taken: Geolocation prediction for social images. *Journal of the Association for Information Science and Technology, 65*(6), 1232–1243.

Liu, J., Li, M., Liu, Q., Lu, H., & Ma, S. (2009). Image annotation via graph learning. *Pattern Recognition, 42*(2), 218–228.

Liu, J., Li, M., Ma, W.-Y., Liu, Q., & Lu, H. (2006). An adaptive graph model for automatic image annotation. In *Proceedings of the 8th ACM international workshop on multimedia information retrieval* (pp. 61–69).

Liu, X., Zhang, J., Zhuo, L., & Yang, Y. (2015). Creating descriptive visual words for tag ranking of compressed social image. In *proceedings of 2015 IEEE international conference on image processing (ICIP)* (pp. 3901–3905).

Longbotham, N., Pacifici, F., Glenn, T., Zare, A., Volpi, M., Tuia, D., & Du, Q. (2012). Multi-modal change detection, application to the detection of flooded areas: outcome of the 2009-2010 data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5*(1), 331–342.

Lu, D., Liu, X. X., & Qian, X. M. (2016). Tag-based image search by social re-ranking. *IEEE Transactions on Multimedia, 18*(8), 1628–1639.

Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Information Retrieval, 14*(2), 178–203.

Ma, W.-Y., & Manjunath, B. S. (1999). Netra: A toolbox for navigating large image databases. *Multimedia Systems, 7*(3), 184–198.

Makita, K., Suzuki, H., Koike, D., Utsuro, T., Kawada, Y., & Fukuhara, T. (2013). Labeling blog posts with wikipedia entries through LDA-based topic modeling of wikipedia. *Journal of Internet Technology, 14*(2), 297–306.

Nguyen, C.-T., Zhan, D.-C., & Zhou, Z.-H. (2013). Multi-modal image annotation with multi-instance multi-label LDA. In *Proceedings of the twenty-third international joint conference on artificial intelligence* (pp. 1558–1564).

Noore, A., Singh, R., & Vatsa, M. (2007). Robust memory-efficient data level information fusion of multi-modal biometric images. *Information Fusion, 8*(4), 337–346.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web. technical report*. CA: Stanford University, Stanford.

Pellegrin, L., Escalante, H. J., Montes-y-Gómez, M., & González, F. A. (2017). Local and global approaches for unsupervised image annotation. *Multimedia Tools and Applications, 76*(15), 16389–16414.

Pentland, A., Picard, R. W., & Sclaroff, S. (1996). Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision, 18*(3), 233–254.

Putthividhya, D., Attias, H. T., & Nagarajan, S. S. (2010a). Supervised topic model for automatic image annotation. In *Proceedings of 2010 IEEE international conference on the acoustics speech and signal processing (ICASSP)* (pp. 1894–1897).

Putthividhya, D., Attias, H. T., & Nagarajan, S. S. (2010b). Topic regression multi--modal latent dirichlet allocation for image annotation. In *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3408–3415).

Qian, X. M., Hua, X. S., Tang, Y. Y., & Mei, T. (2014). Social image tagging with diverse semantics. *IEEE Transactions on Cybernetics, 44*(12), 2493–2508.

Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., Tuytelaars, T., & Van Gool, L. (2005). Modeling scenes with local descriptors and latent aspects. In *Proceedings of tenth IEEE international conference on computer vision* (pp. 883–890).

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1349–1380.

Song, J., Gao, L., Nie, F., Shen, H. T., Yan, Y., & Sebe, N. (2016). Optimized graph learning using partial tags and multiple features for image and video annotation. *IEEE Transactions on Image Processing, 25*(11), 4999–5011.

Srikanth, M., Varner, J., Bowden, M., & Moldovan, D. (2005). Exploiting ontologies for automatic image annotation. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 552–558).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the 2015 IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Tang, J., Shu, X., Qi, G.-J., Li, Z., Wang, M., Yan, S., & Jain, R. (2017). Tri-clustered tensor completion for social-aware image tag refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(8), 1662–1674.

Tousch, A.-M., Herbin, S., & Audibert, J.-Y. (2012). Semantic hierarchies for image annotation: A survey. *Pattern Recognition, 45*(1), 333–345.

Verma, Y., & Jawahar, C. (2017). Image annotation by propagating labels from semantic neighbourhoods. *International Journal of Computer Vision, 121*(1), 126–148.

Virtanen, S., Jia, Y., Klami, A., & Darrell, T. (2012). Factorized multi-modal topic model. In *Proceedings of the 28th conference on uncertainty in artificial intelligence (UAI)* (pp. 843–851).

Wang, C., Jing, F., Zhang, L., & Zhang, H. J. (2007). Content-based image annotation refinement. In *Proceedings of the 2007 IEEE conference on computer vision and pattern recognition* (pp. 1–8).

Wang, C., Shuicheng, Y., Lei, Z., & Zhang, H. J. (2009). Multi-label sparse coding for automatic image annotation. In *Proceedings of the 2009 IEEE conference on computer vision and pattern recognition* (pp. 1643–1650).

Wang, C., Zhang, L., & Zhang, H.-J. (2008). Learning to reduce the semantic gap in web image retrieval and annotation. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 355–362).

Wu, B., Jia, J., Yang, Y., Zhao, P., Tang, J., & Tian, Q. (2017). Inferring emotional tags from social images with user demographics. *IEEE Transactions on Multimedia, 19*(7), 1670–1684.

Xia, Z. Q., Peng, J. Y., Feng, X. Y., & Fan, J. P. (2014). Automatic abstract tag detection for social image tag refinement and enrichment. *Journal of Signal Processing Systems for Signal Image and Video Technology, 74*(1), 5–18.

Xie, Y., Yu, H. M., & Hu, R. (2014). Probabilistic hypergraph based hash codes for social image search. *Journal of Zhejiang University-Science C-Computers & Electronics, 15*(7), 537–550.

Xu, H., Wang, J., Hua, X.-S., & Li, S. (2009). Tag refinement by regularized LDA. In *Proceedings of the 17th ACM international conference on multimedia* (pp. 573–576).

Yan, K., Wang, Y., Liang, D., Huang, T., & Tian, Y. (2016). CNN vs. SIFT for image retrieval: Alternative or complementary. In *Proceedings of the 2016 ACM on multimedia conference* (pp. 407–411).

Yin, Z., Cao, L., Han, J., Zhai, C., & Huang, T. (2011). Geographical topic discovery and comparison. In *proceedings of the 20th international conference on World Wide Web* (pp. 247–256).

Zhai, Z., Liu, B., Xu, H., & Jia, P. (2011). Constrained LDA for grouping product features in opinion mining. In *Proceedings of advances in knowledge discovery and data mining* (pp. 448–459).

Zhang, D., Islam, M. M., & Lu, G. (2012). A review on automatic image annotation techniques. *Pattern Recognition, 45*(1), 346–362.

Zhang, J., Wang, S., & Huang, Q. (2017). Location-based parallel tag completion for geo-tagged social image retrieval. *ACM Transactions on Intelligent Systems and Technology (TIST), 8*(3) Article 38.

Zhang, J., Yang, Y., Tian, Q., Zhuo, L., & Liu, X. (2017). Personalized social image recommendation method based on User-Image-Tag model. *IEEE Transactions on Multimedia, 19*(11), 2439–2449.

Zhang, R., Zhang, Z., Li, M., Ma, W.-Y., & Zhang, H.-J. (2005). A probabilistic semantic model for image annotation and multimodal image retrieval. In *Proceedings of the tenth IEEE international conference on computer vision* (pp. 846–851).

Zhang, X., Zhao, X., Li, Z., Xia, J., Jain, R., & Chao, W. (2013). Social image tagging using graph-based reinforcement on multi-type interrelated objects. *Signal Processing, 93*(8), 2178–2189.

Zheng, L., Wang, S. J., Wang, J. D., & Tian, Q. (2016). Accurate image search with multi-scale contextual evidences. *International Journal of Computer Vision, 120*(1), 1–13.