Learning Probabilistic Kernel from Latent Dirichlet Allocation

Qi Lv[1], Lin Pang[2]*, Xiong Li[2]
1 School of Management and Economics, North China University of Water Resources and Electric Power
Zhengzhou, 450000, China
[e-mail: qiresearch@126.com]
2 National Computer Network Emergency Response Technical Team
Beijing, 100029, China
[e-mail: {panglin,lixiong}@cert.org.cn]
*Corresponding author: Lin Pang

Received January 10, 2015; revised October 23, 2015; revised January 18, 2016; accepted April 7, 2016; published June 30, 2016

Abstract

Measuring the similarity of given samples is a key problem of recognition, clustering, retrieval and related applications. A number of works, e.g. kernel method and metric learning, have been contributed to this problem. The challenge of similarity learning is to find a similarity robust to intra-class variance and simultaneously selective to inter-class characteristic. We observed that, the similarity measure can be improved if the data distribution and hidden semantic information are exploited in a more sophisticated way. In this paper, we propose a similarity learning approach for retrieval and recognition. The approach, termed as LDA-FEK, derives free energy kernel (FEK) from Latent Dirichlet Allocation (LDA). First, it trains LDA and constructs kernel using the parameters and variables of the trained model. Then, the unknown kernel parameters are learned by a discriminative learning approach. The main contributions of the proposed method are twofold: (1) the method is computationally efficient and scalable since the parameters in kernel are determined in a staged way; (2) the method exploits data distribution and semantic level hidden information by means of LDA. To evaluate the performance of LDA-FEK, we apply it for image retrieval over two data sets and for text categorization on four popular data sets. The results show the competitive performance of our method.

Keywords: similarity learning, free energy kernel, LDA, image retrieval, text categorization

This work was supported by NSFC 61403090.

http://dx.doi.org/10.3837/tiis.2016.06.005 **ISSN : 1976-7277**

1. Introduction

In the field of pattern recogntion, similarity measure lies in the focus of classification, clustering, retrieval and related problems. In real world applications, the most challenging work is to find a statisfied feature or similarity measure. For instance, for content based image retrieval, one of the most important components is the similarity measure [1,2,3,4,5]. It is worth noting that, similarity measure can be converted to distance measure and vice versa. Therefore, we do not distinguish the two notations and stick to similarity measure throughout this work. A similarity measure is composed of a feature space and a similarity function over the space, where the feature space is important because it should be robust the intra-class variance and selective to the inter-class attributes. For instance, for image recognition, an ideal feature is robust to illuminance, viewpoint and spatial scale, and selective to semantic content. The similarity function is defined upon the feature space, outputing large value for similar pairs and small value for dissimilar pairs. The most simple measure is the predefined measure, such as L_2 distance [4]. This kind of measures are incapable to adapt data distribution [5] since they have no free parameter to tune. To improve the adaption ability to data distribution, similarity learning methods [3,6,7,8,9] have been proposed. In the perspective of exploiting class label, similarity learning methods are divided into unsupervised learning method and supervised learning method. In the pespective of deriving simiarity measure, learning methods fall to feature space learning and similarity function learning.

Unsupervised similarity learning method seeks to find a feature space or a similarity measure for the training data set, without making use of class label. The methods include factorization methods [10], coding methods [6,7] and probabilistic model based methods [11,12,13,14,15,16]. In these methods, probablistic methods show promsing performance and are received increasing attention. They derive feature mappings [15] or similarity measures [11,12,13] based on the probabilistic models. Thus, they inherite the abilities of probabilistic model, e.g. adaptive to data distribution and capable to infer hidden information. These methods are particularly useful when the class label is missed or is expensive to otain. Supervised similarity learning methods [17,9,18] learn similarity measure by fitting data such that similarity measure outputs large value for sample pair with the same labels and outputs small value for sample pair with distinct labels. Nevertheless, these methods do not fully exploit hidden information and data distribution which could improve the adaption ability and discrimination ability of similarity measure. As a further step, [19,20] exploit class label and probabilistic information simultaneously. However, they still can be further boosted through coupling with more sophisticated probabilstic models.

To exploit data distribution, hidden information and class label for similarity learning, in this work, we propose a similarity learning approach based on latent Dirichlet allocation (LDA) [21] and free energy kernel (FEK) [15], which is referred to as LDA-FEK. The main motivation of this method is to exploit semantic information from LDA and class label which are informative [22,23] for similarity meassure. The method is compatible with bag-of-words represention for given data, where the words can be text words or visual words quantified from image descriptors, and feeds the word histograms to LDA for modelling. Then the free energy kernel is derived based on LDA, essentially being the function of model parameters and variables. To exploit class label, we develop a supervised learning approach for LDA-FEK, which, in technical perspective, tunes the kernel and LDA model to satisfy retreival or recognition performance. The proposed LDA-FEK and its learning method has two main

advantages. First, it can adapt to data distribution, and is able to exploit semantic-level hidden information, i.e. hidden topic and mixture of topic. Second, the supervised learning method can tune the similarity and model according to the retrieval or recognition performance.

The remaining part of this paper is organized as follows. Section 2 reviews the related works of similairty learning. Section 3 derives the LDA-FEK and the learning method. Section 4 applies the proposed approach for retrieval and recognition, and experimentally evaluates it over several popular data sets. Section 5 draws a conclusion.

2. Related Works

A number of works have been contributed to similarity learning [24,25,26,27,28]. We in this work attempt to make a progress on the adaption ability and discrimination abillity of similarity measure. To do so, the posposed method, in technique perspective, should be supervised and probabilistic. In this section, we review supervised similarity learning methods which can be generally categorized into deterministic methods and stochastic methods.

Deterministic similarity learning methods seek to learn similarity measures from the given data set such that the sample pair with the same label has high similarity while the pair with distinct label has low similarity, in a deterministic way. To achieve the purpose, some works [26,29,30] made use of equivalence constraints for pair within the same class, and inequivalence constraints for pair with the different classes. [26] casted the learning problem into a constrained convex optimization problem by minimizing the pairwise distance in the same classes. Discriminative component analysis (DCA) [30] incorporated equivalence constraints with component analysis. Similarly, [31] learned Mahalanobis distance subject to a set of pairwise constraints, i.e., must-links that associate images which must be in the same class and cannot-links that associate samples which must be in different classes.

Also, numerous recent works introduced a variety of techniques for similarity learning Large-margin nearest neighbor (LMNN) [17] coupled with margin maximization criterion. SDPM [32] casted Mahalanobis distance learning to a convex optimization problem. Distance metric learning with eigenvalue optimization (DML-eig) [33] formulated distance learning as an eigenvalue optimization problem. Local distance metric learning (LDML) [9] and [27] learned distance by combining a set of local distance functions. Linear transformation based metric learning (LTML) [18] exploited the flexiblity of linear transformation. Neighborhood component analysis (NCA) [8] cooperated with nearest neighbor criterion. Relevance feedback [34,24], kernel method [18], dimensionality reduction [25], Bayesian inference [35], context [36] and semantic information [22] are also introduced.

Stochastic similarity learning methods derive feautre mapping or similarity measure on the basis of probabilistic generative models. Probability product kernels [12] used the posterior distributions of hidden variables to characterize the samples, and defined the similarity measure as the expected inner product of the hidden variables. [13] used the distributions of observed variables to characterize the samples and used Kullback-Leibler divergence over the distributions to measure the distance between samples. [14] developed a hierarchical probabilistic model to learn representation and similarity. Fisher score (FS) [11] derived explicit feature mapping through considering how a sample affects the model parameters, and defined the Fisher kernel as the weighted inner product of the feature mappings. Free energy score space (FESS) [15] and posterior divergence (PD) [16] extended Fisher score through introducing more informative measures. It is worth noting that, the feature mappings given by these methods are middle level features rather than low level features. These approaches are

able to exploit information from probabilistic generative models. Meanwhile, they still can be improved by tuning the similarity measure to maximize the retrieval or recognition performance. Fisher kernel learning (FKL) [19] learned the kernel parameters by subject to the nearest neighbor criterion. Dsicriminative Fisher kernel learning (DFK) [20] learned Fisher kernel through considering label information. However, the adopted generative model does not fully infer the semantic level information.

In this work, we present an approach LDA-FEK based on the score space methods whose effectivenss has been widely verified [15,16,20]. The advantages of the proposed method are twofold: (1) in comparison with deterministic similarity learning methods, our method exploits data distribution and semantic information in hidden variables; (2) in comparison with other probabilistic learning methods, our method exploits class label in a sophisticated and computationally efficient way.

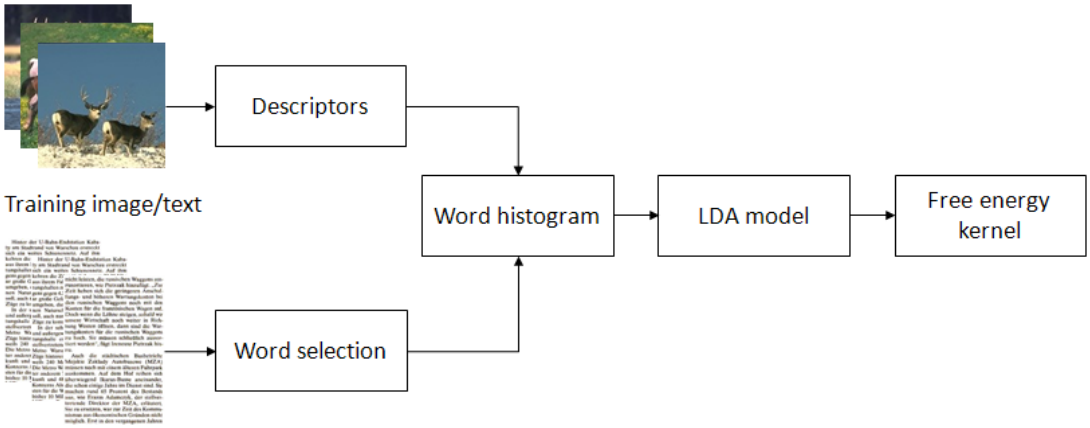


Fig. 1. The framework of our proposed approach LDA-FEK.

3. FESS Kernel from LDA

In this section, we proceed to derive the free energy kernel (FEK) [15] based on Latent Dirichlet Allocation (LDA) [21] and to propose a discriminative learning approach for the kernel. We first represent data as bag-of-words and use LDA to model the distribution for words, due to its effectiveness in text [21] and image modeling [37]. Then, we derive the FEK on the basis of LDA. To boost the discrimination power of the kernel, we propose a discriminative learning method for FEK, which essentially tunes the kernel to satisfy the recognition or retrieval. See **Fig. 1** for the illustration of the proposed method.

3.1. Latent Dirichlet Allocation (LDA)

Given that the samples are represented by bag-of-words which might be quantized from other features, we use LDA [21] to model the distribution of words. LDA was originally proposed for text analysis [21], and was then extended to analyze image by means of bag-of-words representation [37].

LDA is a hierarchical generative model built over a hierarchy of random variables. First, we introduce the mathematical notations. Let $\mathrm{w}=\{w_1,\ldots,w_N\}$ be the document with N words, where $w_n=(w_n^1,\mathrm{L}\ ,w_n^V)^T$ is an indication vector where $w_n^j=1\,(w_n^i=0,\forall i\neq j)$ indicates the

j-th term of all V ones is chosen as the *n*-th word of the document. Let $z_n = (z_n^1, \mathrm{L}, z_n^K)^T$ be the indicative vector for topic, where $z_n^k = 1$ ($z_n^i = 0, \forall i \neq k$) indicates that the *k*-th topic of all K ones is chosen for the *n*-the word.

LDA assumes the following generative process for each document w:

(1) Choose π: $\mathrm{Dir}(\alpha)$. This process can be expressed as a conditional distribution,

$$P(\pi \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \pi_1^{\alpha_1 - 1} \mathrm{L}\ \pi_k^{\alpha_k - 1}$$

(2) For each word w_n of N ones, choose a topic z_n: $\mathrm{Mult}(\pi)$

$$P(z_n \mid \pi) = \prod_{i=1}^{k} \pi_i^{z_{ni}}$$

(3) Choose a word from a conditional multinomial distribution $P(w_n \mid z_n, \beta)$

$$P(w_n \mid z_n, \beta) = \prod_{i=1}^{K} P(w_n \mid \beta_i)^{z_n^i} = \prod_{i=1}^{K} \prod_{j=1}^{V} (\beta_{ij})^{w_n^j z_n^i}$$

Then the joint distribution over the word, topic and mixture of topic can be expressed as,

$$P(\mathrm{w}, \mathrm{z}, \pi \mid \alpha, \beta) = P(\pi \mid \alpha) \prod_{n=1}^{N} P(\mathrm{z}_n \mid \pi) P(w_n \mid \mathrm{z}_n, \beta) \tag{1}$$

3.2. Probabilistic Free Energy Kernel

The standard method to estimate the parameters of probabilistic models is the likelihood maximization method which operates upon the log likelihood of marginal distribution over the observed variable. However, for LDA, it is difficult to obtain the marginal distribution $P(\mathrm{w})$ since the integration of $P(\mathrm{w}, \mathrm{z}, \pi)$ over z, π are intractable [21]. To tackle such problem, it resorts to the variational EM method [38] which instead maximizes the lower bound of log likelihood. To do so, we first construct the approximate posterior distribution of hidden variables as follows,

$$Q(\pi, \mathrm{z} \mid \gamma, \phi) = \mathrm{Q}(\pi \mid \gamma) \prod_{n=1}^{N} Q(z_n \mid \phi_n) = \frac{\Gamma\left(\sum_{i=1}^{K} \gamma_i\right)}{\prod_{i=1}^{K} \Gamma(\gamma_i)} \pi_1^{\gamma_1 - 1} \mathrm{L}\ \pi_K^{\gamma_K - 1} \prod_{n=1}^{N} \prod_{i=1}^{K} \phi_i^{z_{ni}} \tag{2}$$

The approximate posterior distribution $Q(\pi, \mathrm{z} \mid \gamma, \phi)$ takes the same parameterization with the prior distribution $P(\mathrm{z}, \pi)$, but with different parameters. Then the lower bound of the log likelihood function can be derived as,

$$\begin{aligned} \log P(\mathrm{w} \mid \alpha, \beta) &= \log \int \sum_{z} \frac{P(\pi, \mathrm{z}, \mathrm{w} \mid \alpha, \beta) Q(\pi, \mathrm{z} \mid \gamma, \varphi)}{Q(\pi, \mathrm{z} \mid \gamma, \varphi)} d\pi \\ &\geq \mathrm{E}_Q[\log P(\pi, z, w \mid \alpha, \beta) - \log Q(\pi, z \mid \gamma, \varphi)] = L(\theta) \end{aligned} \tag{3}$$

where the inequality is derived by applying Jensen's inequality. The variational lower bound can be further derived as,

$$
\begin{aligned}
L(\theta) &= \log\frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} + \sum_{i=1}^{K}(\alpha_i - 1)\,\mathrm{E}[\log\pi_i] + \sum_{n=1}^{N}\sum_{i=1}^{K}\mathrm{E}[z_{ni}\log\pi_i] \\
&\quad - \log\frac{\Gamma\left(\sum_{i=1}^{K}\gamma_i\right)}{\prod_{i=1}^{K}\Gamma(\gamma_i)} - \sum_{i=1}^{K}(\gamma_i - 1)\,\mathrm{E}[\log\pi_i] - \sum_{n=1}^{N}\sum_{i=1}^{K}\mathrm{E}[z_{ni}\log\varphi_i] + \sum_{n=1}^{N}\sum_{i=1}^{K}\sum_{j=1}^{V} w_{nj}\mathrm{E}[z_{ni}]\log\beta_{ij} \\
&= \log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) - \log\Gamma\left(\sum_{i=1}^{K}\gamma_i\right) + \sum_{i=1}^{K}\left(\Gamma(\alpha_i) - \Gamma(\gamma_i)\right) \\
&\quad + \sum_{i=1}^{K}(\alpha_i - \gamma_i)\,\mathrm{E}[\log\pi_i] + \sum_{n=1}^{N}\sum_{i=1}^{K}\varphi_{ni}\,\mathrm{E}[\log\pi_i - \log\varphi_{ni}] + \sum_{n=1}^{N}\sum_{i=1}^{K}\sum_{j=1}^{V} w_{nj}\varphi_{ni}\log\beta_{ij}
\end{aligned}
\tag{4}
$$

The learning procedure of LDA is the iterative maximization with respect to model parameters (E-step) and posterior distributions (M-step). The details can be found in [21].

Having the variational lower bound $L(\theta)$ of the log likelihood function $\log P(\mathrm{w})$, the score functions of FESS are the fractions of lower bound,

$$
\begin{aligned}
&\phi_\pi(\mathrm{w};\theta):\ \log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) - \log\Gamma\left(\sum_{i=1}^{K}\gamma_i\right),\ \Gamma(\alpha_i) - \Gamma(\gamma_i),\ (\alpha_i - \gamma_i)\mathrm{E}[\log\pi_i],\ \ \forall i \\
&\phi_z(\mathrm{w};\theta):\ \sum_{n=1}^{N} z_{ni}\,\mathrm{E}[\log\pi_i - \log\varphi_i],\ \ \forall i \\
&\phi_w(\mathrm{w};\theta):\ \sum_{n=1}^{N} w_{nj}\varphi_{ni}\log\beta_{ij},\ \ \forall i, j
\end{aligned}
$$

The complete FESS score function is the combination of the above fractions,

$$
\phi(\mathrm{w};\theta) = \left(\phi_\pi(\mathrm{w};\theta)^T, \phi_z(\mathrm{w};\theta)^T, \phi_w(\mathrm{w};\theta)^T\right)^T \tag{5}
$$

These score functions are the expectation of the functions of the random variables and model parameters, where the hidden variables allow FESS kernel to exploit the hidden information, and model parameters allow it to adapt to data distribution.

The free energy kernel then can be defined as,

$$
K(\mathrm{w}_i, \mathrm{w}_j; U, \theta) = \phi(\mathrm{w}_i;\theta)^T U \phi(\mathrm{w}_j;\theta) \tag{6}
$$

where U is the weight matrix and, in the followin section, will be determined by fitting to data in the learning procedure.

3.3. Learning LDA based Free Energy Kernel

Let y_i be the label vector for a specific sample w_i. We consider the criterion that a pair of samples takes high similarity if they have the same label, and takes low similarity if they have the distinct label. The objective function can be expressed as,

$$
J(U, \theta) = \sum_i \sum_{j \neq i} s(\mathbf{y}_i, \mathbf{y}_j) K(\mathrm{w}_i, \mathrm{w}_j; U, \theta) \tag{7}
$$

The label similarity function $s(\mathbf{y}_i, \mathbf{y}_j)$ tends to have larger value if they have more common labels, and will be specified according to applications. For simplification, we assume that the weight maxtrix is diagonal U=diag(u_1, L, u_D), where u_d weights the importance of ϕ_d to the similarity. Note that, $K(\mathrm{w}_i, \mathrm{w}_j; U, \theta) = \sum_d \phi(\mathrm{w}_i;\theta)_d u_d \phi(\mathrm{w}_j;\theta)_d$, where u_d weights the importance of the *d*-th component of the feature mapping ϕ. It encourages those components with satisfied discrimination ability, and inhibits those components with unsatisfied

discrimination ability. Thus, it plays the role of feature selection in the kernel-based similarity measure. On the other hand, this weight provides external adaption ability to data distribution or intrinsic pattern beside θ. In the case of $u_d > 0$, the above equation can be formulated as $\sum_d \sqrt{u_d}\phi(\mathrm{w}_i;\theta)_d \cdot \sqrt{u_d}\phi(\mathrm{w}_j;\theta)_d$, where $\sqrt{u_d}$ implements a linear mapping for ϕ.

Due the computational cost, we consider a *stage* learning procedure, i.e., first learn model parameter θ and then learn weight matrix U. The unknown weight matrix can be determined through maximizing $J(U)$ with respect to U using gradient descent method,

$$\frac{\partial J(U,\theta)}{\partial u_d} = \sum_i \sum_{j \neq i} s(\mathbf{y}_i, \mathbf{y}_j)\phi(\mathrm{w}_i;\theta)^T \phi(\mathrm{w}_j;\theta) \tag{8}$$

The learning procedure of the proposed approach is the iteration of Eq.(8), which is summarized in **Algorithm 1.** The solution of gradient descend might go to local minima. To relieve such a problem, a pre-train strategy is adopted in our implementation. First, U is initialized as an identical matrix, i.e. $u_d = 1$, where all components of the feature mapping ϕ are assumed to be equally important at start. Second, run Algorithm 1 on a subset of training set, giving the solution $U^{(0)}$. (3) Run **Algorithm 1** on the whole training set, where $U^{(0)} = U^{(1)}$ is used as the initial value.

Algorithm 1 Learning LDA-FEK

1: Input: training set $\{(\mathrm{w}_i, y_i)\}_{i=1}^N$; iteration number T; learning rate $\gamma > 0$

2: Initialize parameters $U^{(0)}$ through pre-train

3: Train LDA $\hat{\theta}$ using variational EM algorithm

4: For $t = 1$ to T do

5: $U^{(t)} \leftarrow U^{(t-1)} - \gamma \frac{\partial J(U;\hat{\theta})}{\partial U}$

6: End for

7: Output: $U^{(T)}$

The learned LDA-FEK can be embeded to any kernel based classifier. In classification procedure, the kernel similarity of w_i and w_j can be computed using **Algorithm 2.**

Algorithm 2 Computing LDA-FEK similarity

1: Input: a pair of documents $\mathrm{w}_i, \mathrm{w}_j$

2: Compute posterior $Q(\pi, z \mid \mathrm{w}_i)$ (variational inference [21] or Gibbs sampling [39])

3: Compute posterior $Q(\pi, z \mid \mathrm{w}_j)$ (variational inference [21] or Gibbs sampling [39])

4: compute the free energy kernel similarity using Eq.(5) and (6)

5: Output: $K(\mathrm{w}_i, \mathrm{w}_j)$

Here we use the text categorization experiment in Section 4.2 with Pascal dataset to demonstrate the convergence of our method. We visulize the convergence procedure of

Algorithm 1 in **Fig. 2**. As shown in **Fig. 2**, **Algorithm 1** reaches convergence within about 50 iterations, which is quite efficient. According to our experiments, this learning procedure is about 8-10x faster than the algorithm with the pretraining step removed, without lossing the retrieval of categorization performance. For **Algorithm 2**, it is highly efficient scince all steps require no iteration.

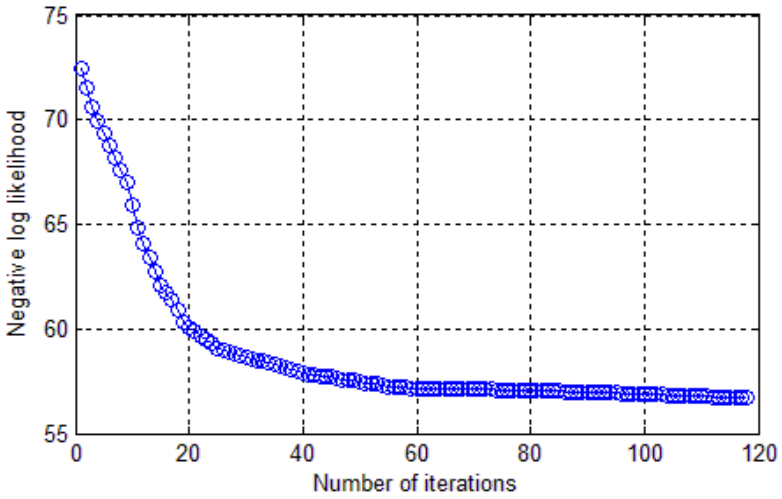


Fig. 2. The negative log likelihood as a function of the training step

4. Experiments

In this section, we apply our Latent Dirichlet Allocation based free energy kernel (LDA-FEK) for image retrieval and text categorization. The proposed method will be compared with several state-of-the-art methods on Corel5K [40] and MIRFLICKR [41] datasets for image retrieval and on 20 News, Sentiment, Reuters and Pascal datasets for text categroization.

4.1 Image Retrieval

The last decade has seen the increasing popularity of digital images. How to search images according to user input from a large gallery has been an important and challenging problem [5,1,2]. We are interested in the problem of retrieval by image [5,34,1], where is referred to as content based image retrieval (CBIR).

The framework of applying LDA-FEK for image retrieval is illustrated in **Fig. 3**. First, we quantized image descriptors to visual words, and train LDA-FEK on the training set for a chosen number of topics. In the retrieval procedure, for a query image, first represent it as a word histogram, and infer its corresponding hidden variables. Its similarity with respect to a candidate image is computed using **Algorithm 2**. For computational effectiveness, we use Gibbs sampling [39] to fit the LDA model.

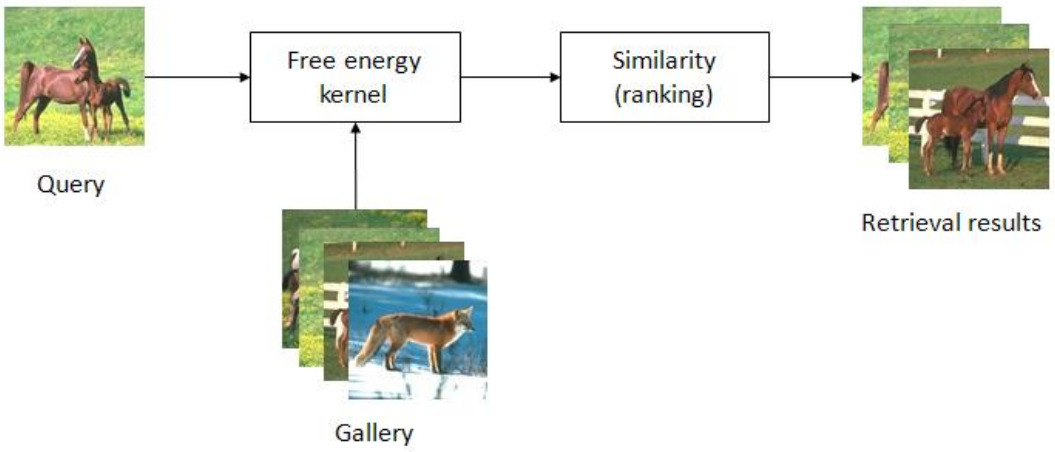


Fig. 3. The framework of image retrieval via LDA-FEK.

4.1.1 Feature Representation and Performance Measure

Image feature is an important component for retrieval systems. It is expected to capture useful information and discard intra-class variance. In this experiment, we follow the suggestions in [42] and use four types of color SIFT descriptors as the low level features, i.e., OpponentSIFT, rgSIFT, C-SIFT and RGB-SIFT. These descriptors are extracted from the image patches sampled from dense grid and Harris-Laplace interest point, with spatial pyramid followed.

Let $\mathrm{y}_i = (y_{1i}, \mathrm{L}\ , y_{Ci})^T$ be the label vector for image w_i, where $y_{ci} = 1$ if w_i belongs to the label c and $y_{ci} = 0$ otherwise. We choose the label similarity $s(\mathbf{y}_i, \mathbf{y}_j)$ as follows:

$$s(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{y}_i^T \mathbf{y}_j \tag{9}$$

Then the direction of gradient descent is,

$$\frac{\partial J(U,\theta)}{\partial u_d} = \sum_i \sum_{j \neq i} \mathbf{y}_i^T \mathbf{y}_j \phi(\mathrm{w}_i)^T \phi(\mathrm{w}_j)$$

The learning algorithm is obtained by embedding it to **Algorithm 1**.

In this experiment, we evaluate the retrieval performance using leave-one-out manner [33,3,4]. First, choose a query image from the test set. Second, find the similar images from the candidate set according to the LDA-FEK similarity measure. We use mean average precision (MAP) which is the summarization of the precision-recall curve, to measure the retrieval performance. The precision is defined as the percentage of returned images that contain the same label with the query image.

Let k be the rank, the precision at cut-off k can be computed as:

$$P(k) = \frac{|\{\text{relevant retrieved images of rank } k \text{ or less}\}|}{k}$$

The average precision (AP) is the averaging of the precision for relevant returned images, that is, $\mathrm{AP} = \sum_{k=1}^{K} P(k) \cdot \mathrm{rel}(k) / K$,where K is the number of relevant images, $\mathrm{rel}(k)$ indicates whether the image at the rank k is relevant. Averaging AP across all the query images gives,

$$\mathrm{MAP} = \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{AP}(q)$$

where Q is the number of query images. Note that, since the datasets generally contain mutiple categoires and each category compries distinct number of sample, the MAP can low. Further, to evaulate the variance of the retrival results, we also report the standard deviation of AP across queries.

4.1.2 Experimental Results on Corel5K dataset

First we perform the experiment on Corel5K dataset [40]. The Corel5k dataset is a subset of Corel Photo Gallery, comprising 50 categories, e.g. beach, tile, wave, tigers, France, bears, autumn, and tropical plants, where each category contains 100 images. It has a vocabulary of 371 words. The sizes of the images are normlized to 192×128 or 128×192, see **Fig. 4** for examples. In this experiment, we randomly choose 70% samples as the training set and remain the rest as the test set, where the training set is used to learn PLSA-FK and the test set is used to evaluate the performance. For all compared approaches, we compute the average precision (AP) for each category over the top 20 retrieved images.

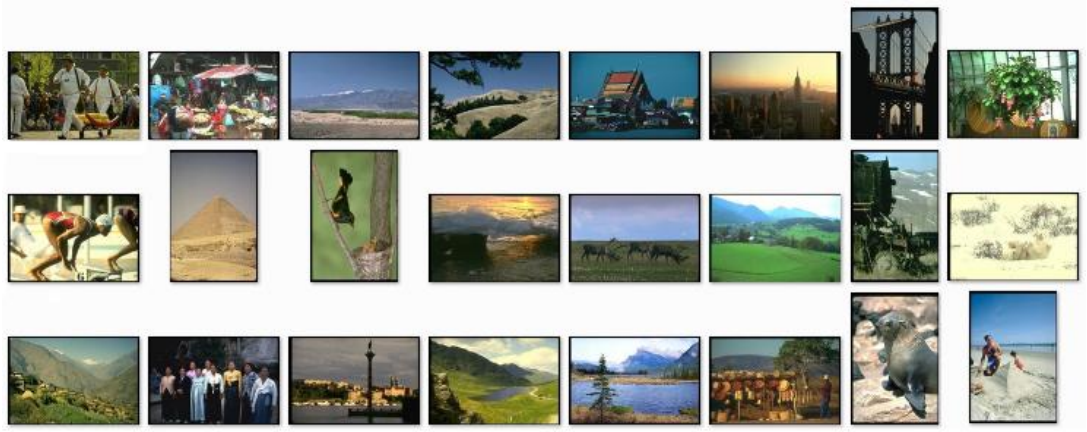


Fig. 4. Sample images of Corel5K dataset

We compare the proposed approach LDA-FEK with several state-of-the-art approaches: SDPM [32], DML-eig. [33], large margin nearest neighbor (LMNN) [17], local distance metric learning (LDML) [9], free energy score space (FESS) [15] and Fisher kernel learning (FKL) [19]. For FESS, we derive the feature mapping from LDA and defined the kernel similarity as the inner product, which is similar with Eq. (6) except that the weight matrix is removed. For FKL, we couple it with Gaussian mixture model. FKL is originally designed for classification. Here we extend it from retrieval by replacing its label similarity component with Eq. (9). For our approach, the number of topics of LDA is set to $K=100$ through cross validation on the training set. For the compared approaches, we implement FESS and DFK and follow the authors' settings, and refer to the results of other from literatures.

The experimental results are reported in **Table 1**. It can be found that, DML-eig and LMNN show similar performance. Meanwhile, SDPM outperforms DML-eig and LMNN. The underlying reason is that SDPM reaches a good solution by means of convex optimization. Probabilistic similarity learning approaches, FKL and FESS, show competitive performance because they exploit image distribution and topic information. The proposed method LDA-FEK, in most cases, achieves the best performance against the compared approaches. Specifically, LDA-FEK outperforms FESS by about 1.8%. This improvement is credited to the label information which FESS does not utilize. Also, LDA-FEK outperforms FKL about

1.1%, because LDA is able to infer the semantic level hidden information much better that GMM. Considering the standard deviation of AP cross all queries, these results demonstrate the effectiveness of the proposed method in image retrieval.

Table 1. Retrieval performance over Corel5K dataset

Algorithm	MAP (mean average precision)	standard deviation (AP)
SDPM [32]	0.315	0.008
DML-eig [33]	0.309	0.012
LMNN [17]	0.310	0.010
LDML [9]	0.319	0.013
FESS [15]	0.320	0.009
FKL [19]	0.327	0.011
LDA-FEK (ours)	**0.338**	0.010

Moreover, we note that the query image could be highly various, typically with distinct noise. Here we evaluate the robustness of our approach to noise. We simulate the noisy query images by adding Gaussian noise to each pixel, where the mean value is set to the original value and the bandwidth (sigma) is set to 20, 30 and 40 respectively. The noise test is highly challenging. The MAP of our approach LDA-FEK goes down to 0.301, 0.287 and 0.236 respectively. It can be found that, the decreasing trend of MAP along the noise level is getting sharp, which suggests that, the proposed approach works when the bandwidth < 20, and suddenly gets worse when bandwidth > 20. This fact also implies that, denoise techniques might be helpful to image retrieval. Further, we evaluate the performance of our method when the number of training examples varies. 20%, 30%, 40, 60%, 80% examples are sampled from the training set and are used to train the model. The results of our method and the best compared method FLK (**Table 1**) are reported in **Fig. 5**. It can be found that, our method achieves an improvement of about 4% over FKL when the number of training examples is small, which suggests that, as a benefit of exploiting generative model, our method can be applied to those situations with only few training examples.

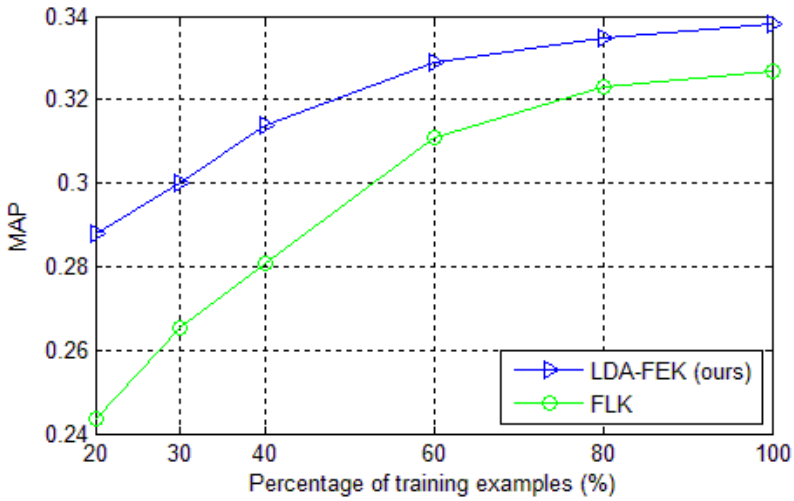


Fig. 5. MAP as a function of the percentage of training examples

4.1.3 Experimental Results on MIRFLICKR dataset

This experiment evaluates the performance of our method on large dataset. In real world applications, the dataset is usually very large, with great intra-class variance. These conditions require that the similarity measure is scalable, and could capture semantic level information against the intra-class variance. The MIRFLICKR dataset [41], collected from Flickr which is an online photo-sharing website, is used for experiment. It contains 25,000 samples with high-resolution images and corresponding text annotations. The size of the images are normalized to MAX (WIDTH, HEIGHT) =500. Some sample images are shown in **Fig. 6**. For comparison, we follow the following experimental scheme. The dataset is split into two parts, 15,000 images for training and the rest 10,000 images for test. 1,000 images are randomly chosen from the test dataset as query images and the rest 24,000 images are remained as the gallery. In the gallery, 15,000 images are with text annotations.

Fig. 6. Sample images from MIRFlickr dataset

The proposed method LDA-FEK is compared with several state-of-the-art methods: nonnegative matrix factorization (NNMF) [10], large margin nearest neighbor (LMNN) [17], linear transformation based metric learning (LTML) [18], free energy score space (FESS) [15] and Fisher kernel learning (FKL) [19]. NNMF is a state-of-the-art method on the basis of matrix factorization. LMNN is a supervised distance learning method under the large margin criterion. FESS and FKL are probabilistic similarity learning methods closely related to our method. For all compared methods, we used the authors' suggested settings. For our method, the number of topics in LDA is set to K=160 according to cross validation.

Table 2. The retrieval performance on MIRFLICKR dataset.

Algorithm	MAP (mean averge precision)	standard deviation (AP)
NNMF [10]	0.583	0.015
LMNN [17]	0.586	0.020
LTML [18]	0.597	0.011
FESS [15]	0.592	0.017
FKL [19]	0.601	0.014
LDA-FEK (ours)	0.613	0.013

The experimental results, i.e., MAP and standard deviation of AP, are reported in **Table 2**. It can be found that, FESS and FKL outperform NNMF and LMNN. An important reason is

that, compared with NNMF and LMNN, FESS and FKL exploit the data distribution and semantic information in a more sophisticated way. Also, LTML shows competitive performance with FESS, which benefits from the flexibility of linear transformation. Further, LDA-FEK shows superiority over FESS, about 2.1%, with the competitive standard deviation. The reason is that, LDA-FEK exploits class label through tuning similarity measure with respect to retrieval performance. Also, LDA-FEK outperforms FKL about 1.2%, because it benefits from the semantic level information inferred by LDA. Also, as did in Section 4.1.2, we evaluate the performance of our method over varying number of training examples. 20%, 30%, 40, 60%, 80% examples are sampled to form the training set. The results of our method and the state-of-the-art method LTML are reported in **Fig. 7**. Our method again shows superiority (up to 6%) than LTML when the number of training examples is small. The primary reason is that our method benefits from Bayes inference.

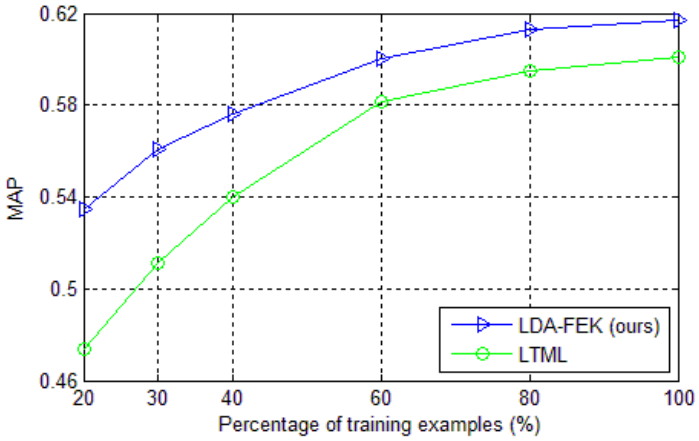


Fig. 7. MAP as a function of the percentage of training examples

4.2 Text Categorization

The proposed method LDA-FEK can also be applied to text categorization since LDA is orignally designed for text analysis. The framework of applying our LDA-FEK for text categorization is illustrated in **Fig. 8**. Note that, the kernel is learned beforehand and then isembedded into the classifier for recognition.

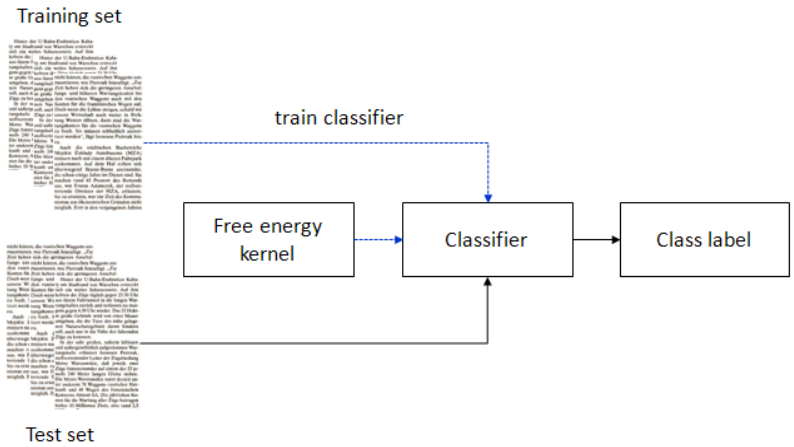


Fig. 8. The framework of text categorization via LDA-FEK.

In this experiment, the label similarity function in Eq.(7) is chosen as follows,

$$s(y_i, y_j) = 2\mathrm{I}(y_i = y_j) - 1$$

It outputs 1 if y_i equals y_j and -1 otherwise. Then the direction of gradient descent is,

$$\frac{\partial J(U,\theta)}{\partial u_d} = \sum_i \sum_{j \neq i} (2\mathrm{I}(y_i = y_j) - 1)\phi(\mathrm{w}_i)^T \phi(\mathrm{w}_j)$$

The learning algorithm is obtained by substituting the above equation into **Algorithm 1**. For computational effectiveness, we use Gibbs sampling [39] to fit LDA model.

4.2.1 Data sets and Feature Representation

The 20 Newsgroups corpus contains about 20,000 messages from 20 distinct newsgroups. Following the previous works [43], we construct three recognition tasks: (1) Comp: comp.sys.ibm.pc.hardware vs. comp.sys.mac.hardware; (2) Sci: sci.electronics vs. sci.med; (3) Talk: talk.politics.guns vs. talk.politics.mideast. Each message is represented by bag-of-words. For each problem, we selected 1800 examples balanced between two labels.

The Reuters Corpus contains over 800,000 newswire stories [44]. Each article contains one or more labels describing its general topic, industry and region. We created the following binary recognition tasks from the labeled documents: (1) Insurance: Life vs. Non-Life; (2) Business: Banking vs. Financial; (3) Retail: Specialist Stores vs. Mixed Retail. These tasks involve similar categories so they are hard to recognize. For each problem, we selected 2000 samples using a bag-of-words representation. Each problem contains a balanced mixture of samples from each label.

The sentiment multi-domain data set of [21] consists of product reviews from 7 Amazon domains (apparel, book, dvd, electronics, kitchen, music, video). The goal in each domain is to classify a product review as either positive or negative. Feature extraction creates unigram and bigram features using counts following [21]. For the apparel domain we used all 1940 samples and for all other domains we used 2000 samples. Each problem contains a balanced mixture of example labels.

The PASCAL large scale learning challenge workshop provided several large scale binary data sets. We selected the NLP task which is a Webspam filtering problem. Each sample is the text from a web page. The task is to classify a webpage as either spam or ham. We used the default format provided by the workshop and selected 2000 samples.

4.2.2 Experimental Results

In this experiment, we compare the performance of the LDA-FEK with LDA-NB, LDA-FESS [15] and diagonalized CW [43]. LDA-NB trains a LDA for each category and uses the maxiumum a posteriori rule for decision. LDA-FESS derives FESS feature based on LDA and delivers to SVMs for categorization. Diagonalized CW, confidence weighted learning method, is a state-of-the-art method of text categorization. The data sets, feature representation and other details are list in the above section. Linear SVM is used as the classifier with our LDA-FEK embedded.

Table 3. The text categorization performance on 4 data sets.

Data set	Task	SVM	CW	LDA-NB	LDA-FESS	LDA-FEK
Sentiment	Apparel	13.92	12.53	12.93	13.24	**11.87**
	Books	18.25	**16.90**	18.10	17.05	17.24
	DVD	19.60	**17.45**	20.13	18.86	18.21
	Electronics	16.25	14.95	17.02	**14.91**	15.02
	Kitchen	15.50	13.75	14.90	14.38	**13.25**
	Music	18.25	17.15	18.42	**17.09**	17.20
	Video	18.80	21.75	21.98	18.92	**18.12**
Reuters	Retail	12.90	10.55	14.65	12.24	**9.95**
	Business	**15.60**	16.35	19.74	15.82	15.70
	Insurance	9.75	**8.20**	11.30	9.47	9.16
20 News	Comp	7.67	6.69	5.65	6.97	**5.58**
	Sci	3.86	2.44	**1.82**	3.50	2.04
	Talk	1.24	**0.86**	0.98	1.33	0.95
Pascal	Webspam	3.85	3.55	15.26	4.19	**3.31**

The average catagorization errors on all four data sets are reported in **Table 3**, where the results of SVM and CW are referred in [43]. It can be found that, LDA-FESS outperform SVM on 10 tasks and also outperforms LDA-NB on 10 tasks, which consistents with the previous works [15]. Futher, LDA based methods show competitive performance with SVM and CW, which suggests that the semantic level information, i.e., topic and mixture of topic, inferred by LDA is informative for text categorization. Moreover, our LDA-FEK shows superiority over LDA-FESS on 11 tasks, which suggests that the discriminative learning which exploits label information is effective. We also find that, these algorithms show perference to data sets. LDA-FS works particularly well on 20 News while CV works well on Pascal. As a potential reason, the basis introduced in algorithms would induce the perference to data sets. Further, as did in the above experiments, the performance of our method over varying number of training examples is evaluated. 20%, 30%, 40, 60%, 80% examples are sampled from Pascal dataset to form the training set. The results of our method and the best compared method CW (**Table 3**) are reported in **Fig. 9**. Our method achieves an improvement up to 8% against CW, when the training set is small, which verifies the advantage of exploiting generative information in similarity learning.

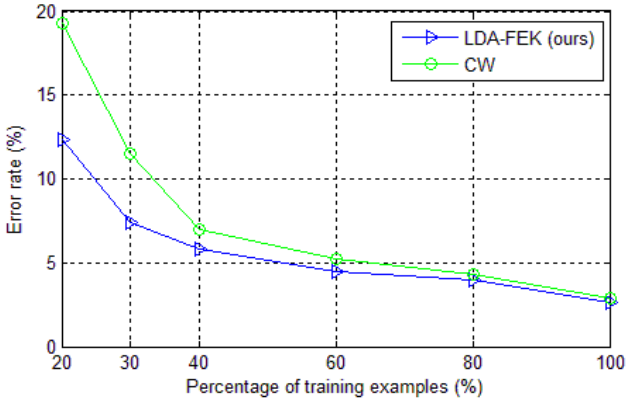


Fig. 9. Error rate as a function of the percentage of training examples

5. Conclusions

In this paper, we proposed a similarity learning approach on the basis of LDA which is able to discover the topic and mixture of topic hidden in data. The similarity, i.e., free energy kernel, is a function over the parameters and variables of LDA. Thus, it inherited the capability of data adapation and semantic information inferred from LDA for retrieval and recognition. The semantic information given by LDA are the topic and scene in image, or topic and mixture of topic in text. Although free energy score space (FESS) is a state-of-the-art feature, it is further boosted in our framework by means of intrducing the additional flexiability (matirx U) by which the method could adapt to both data distribution and tune for the performance much better. Moreover, FESS itself can also tune for the retrieval or categroization performance through joint optimization. The proposed method LDA-FEK is applied to both image retrieval and text categorization. The results suggest that proposed method is comptitive to other state-of-the-art methods in performance and is scalable in data set. However, condering the fact that SIFT base features are good at capturing local texture information instead of global shape information, the proposed method can be potentially improved by introducing some complementary features, e.g. global features.

References

[1] F. Faria, A. Veloso, H. Almeida, E. Valle, R. Torres, M. Gonc¸alves and W. Meira Jr, "Learning to rank for content-based image retrieval," in *Proc. of ACM Conference on Multimedia Information Retrieval*, pp. 285–294, 2010. Article (CrossRef Link)

[2] M. Arevalillo-Herr´aez, F. Ferri and J. Domingo, "A naive relevance feedback model for content-based image retrieval using multiple similarity measures," *Pattern Recognition*, 43(3):619–629, 2010. Article (CrossRef Link)

[3] L. Yang, R. Jin, L. Mummert, R. Sukthankar, A. Goode, B. Zheng, S. Hoi and M. Satya-narayanan, "A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):30–44, 2010. Article (CrossRef Link)

[4] S. Hoi, W. Liu and S. Chang, "Semi-supervised distance metric learning for collaborative image retrieval and clustering," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 6(3), 2010. Article (CrossRef Link)

[5] A. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000. Article (CrossRef Link)

[6] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3360–3367, 2010. Article (CrossRef Link)

[7] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*. 290(5500):2323–2326, 2000. Article (CrossRef Link)

[8] J. Goldberger, S. Roweis, G. Hinton and R. Salakhutdinov, "Neighborhood components analysis," In *NIPS*, 2004. Article (CrossRef Link)

[9] L. Yang, R. Jin, R. Sukthankar and Y. Liu, "An efficient algorithm for local distance metric learning," in *Proc. of the National Conference on Artificial Intelligence*, 2006. Article (CrossRef Link)

[10] J.C. Caicedo, J. BenAbdallah, F.A. González and O. Nasraoui. "Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization," *Neurocomputing*. 76(1), pp. 50-60, 2012. Article (CrossRef Link)

[11] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," In *NIPS*, 1998. Article (CrossRef Link)
[12] T. Jebara, R. Kondor and A. Howard, "Probability product kernels," *Journal of Machine Learning Research*, 5:819-844, 2004. Article (CrossRef Link)
[13] N. Vasconcelos. "On the efficient evaluation of probabilistic similarity functions for image retrieval," *IEEE Transactions on Information Theory*. 50(7):1482-1496, 2004. Article (CrossRef Link)
[14] C. Schmid, "Constructing models for content-based image retrieval," in *Proc. of CVPR* 2001. Article (CrossRef Link)
[15] A Perina, M. Cristani, U. Castellani, V. Murino and N. Jojic. "Free energy score spaces: using generative information in discriminative classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. Article (CrossRef Link)
[16] X. Li, T.S. Lee and Y. Liu. "Hybrid generative-discriminative classification using posterior divergence," In *CVPR,* 2011. Article (CrossRef Link)
[17] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*. 10:207–244, 2009. Article (CrossRef Link)
[18] P. Jain, B. Kulis, J. Davis and I. Dhillon, "Metric and kernel learning using a linear transformation," *The Journal of Machine Learning Research* 13:519–547, 2012. Article (CrossRef Link)
[19] L.V. der Maaten, "Learning discriminative fisher kernels," in *Proc. of ICML*, pp. 217–224, 2011. Article (CrossRef Link)
[20] B. Wang, X. Li and Y. Liu, "Learning discriminative Fisher kernel for image retrieval," *KSII Transaction on Internet and Information System*, 7(3), 2013. Article (CrossRef Link)
[21] J. Blitzer, M. Dredze and F. Pereira. "Biographies, Bollywood, boom-boxes and blenders, Domain adaptation for sentiment classification," in *Proc. of ACL*, 2007. Article (CrossRef Link)
[22] J. Yu, D. Tao, J. Lic and J. Cheng, "Semantic preserving distance metric learning and applications," *Information Sciences*, 281:674-686, 2014. Article (CrossRef Link)
[23] B. Liu, M. Wang, R. Hong, Z.J. Zha and X.S. Hua. "Joint learning of labels and distance metric," *IEEE Transactions on Systems, Man and Cybernetics*, 40(3):973-978, 2010. Article (CrossRef Link)
[24] J. Su, W. Huang, P. Yu and V. Tseng, "Efficient relevance feedback for content-based image retrieval by mining user navigation patterns," *IEEE Transactions on Knowledge and Data Engineering*, 23(3):360–372, 2011. Article (CrossRef Link)
[25] H. Cai, K. Mikolajczyk and J. Matas, "Learning linear discriminant projections for dimensionality reduction of image descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 33(2):338–352, 2011. Article (CrossRef Link)
[26] E. Xing, A. Ng, M. Jordan and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Proc. of NIPS*, pp.505–512, 2002. Article (CrossRef Link)
[27] A. Frome, Y. Singer and J. Malik, "Image retrieval and classification using local distance functions," in *Proc. of NIPS*, 2007. Article (CrossRef Link)
[29] A. Bar-Hillel, T. Hertz, N. Shental and D. Weinshall, "Learning distance functions using equivalence relations," in *Proc. of Machine Learning-International Workshop Then Conference*, 2003. Article (CrossRef Link)
[30] S. Hoi, W. Liu, M. Lyu and W. Ma, "Learning distance metrics with contextual constraints for image retrieval," in *Proc. of CVPR*, pp. 2072–2078, 2006. Article (CrossRef Link)
[31] S. Xiang, F. Nie and C. Zhang, "Learning a mahalanobis distance metric for data clustering and classification," *Pattern Recognition* 41(12):3600–3612, 2008. Article (CrossRef Link)
[32] J. Kim, C. Shen and L. Wang, "A scalable algorithm for learning a Mahalanobis distance metric," in *Proc. of ACCV*, 2010. Article (CrossRef Link)
[33] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization," *The Journal of Machine Learning Research* 13:1–26, 2012. Article (CrossRef Link)

[34] S. Hoi, M. Lyu and R. Jin, "A unified log-based relevance feedback scheme for image retrieval," *IEEE Transactions on Knowledge and Data Engineering*, 18(4):509–524, 2006. Article (CrossRef Link)

[35] L. Yang, R. Jin and R. Sukthankar, "Bayesian active distance metric learning," *arXiv preprint arXiv*, 1206.5283, 2012. Article (CrossRef Link)

[36] H. Becker, M. Naaman and L. Gravano, "Learning similarity metrics for event identification in social media," in *Proc. of ACM international conference on Web search and data mining*, pp. 291–300, 2010. Article (CrossRef Link)

[37] L. Fei-Fei and P. Perona. "A bayesian hierarchical model for learning natural scene categories," in *Proc. of CVPR*, 2005. Article (CrossRef Link)

[38] M. Jordan, Z. Ghahramani, T. Jaakkola, and S. Lawrence. "Introduction to variational methods for graphical models," *Machine Learning*, 37:183-233, 1999. Article (CrossRef Link)

[39] T. Griffiths and M. Steyvers. "Finding scientific topics," in *Proc. of the National Academy of Sciences*. 101(Suppl 1):5228-5235, 2004. Article (CrossRef Link)

[40] P. Duygulu, K. Barnard, N. de Freitas and D. Forsyth., "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. of ECCV*, 2002. Article (CrossRef Link)

[41] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. of ACM International Conference on Multimedia Information Retrieval*, 2008. Article (CrossRef Link)

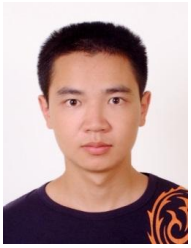
[42] K. Van De Sande, T. Gevers and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 32(9) :1582–1596, 2010. Article (CrossRef Link)

[43] K. Crammer, M. Dredze and F. Pereira. "Confidence-weighted linear classification for text categorization," *Journal of Machine Learning Research*, 13:1891-1926, 2012. Article (CrossRef Link)

[44] D. Lewis, Y. Yang, T. Rose and F. Li. Rcv1, "A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, 5:361–397, 2004. Article (CrossRef Link)

Qi Lv received the BS and MS degrees in Flight Vehicle Propulsion Engineering from Nanjing University of Aeronautics and Astronautics, China, in 2002 and 2005 respectively, and PhD degree in mathematics from Zhengzhou University, China, in 2010. His research interests include pattern recognition, statistics and Internet public opinion.

Lin Pang received her PhD degree in Computer Science and Technology from Institute of Computing Technology, Chinese Academy of Sciences in 2012. She is currently an engineer in National Computer Network Emergency Response Technical Team of China. Her research interests include machine learning, multimedia analysis and retrieval.



Xiong Li received the PhD degree in pattern recognition and intelligence system from Shanghai Jiao Tong University, China, in 2013. He is currently a senior engineer in National Computer Network Emergency Response Technical Team, China. His research interests include hybrid generative discriminative learning and probabilistic graphical model.