Heterogeneous Metric Learning for Cross-Modal Multimedia Retrieval

Jun Deng[1,2], Liang Du[1,2], and Yi-Dong Shen[1]

[1]State Key Laboratory of Computer Science, Institute of Software
Chinese Academy of Sciences, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
{dengj,duliang,ydshen}@ios.ac.cn

Abstract. Due to the massive explosion of multimedia content on the web, users demand a new type of information retrieval, called cross-modal multimedia retrieval where users submit queries of one media type and get results of various other media types. Performing effective retrieval of heterogeneous multimedia content brings new challenges. One essential aspect of these challenges is to learn a *heterogeneous metric* between different types of multimedia objects. In this paper, we propose a Bayesian personalized ranking based heterogeneous metric learning (BPRHML) algorithm, which optimizes for correctly ranking the retrieval results. It uses pairwise preference constraints as training data and explicitly optimizes for preserving these constraints. To further encouraging the smoothness of learning results, we integrate graph regularization with Bayesian personalized ranking. The experimental results on two publicly available datasets show the effectiveness of our method.

Keywords: Metric Learning, Heterogeneous Spaces, Multimedia.

1 Introduction

With the explosive accumulation of multimedia content on the web, cross-modal multimedia retrieval attracts much attention of industry and academia. Nowadays, the prevailing tools for retrieving multimedia content are still single-media based, e.g., search engines such as google or bing. In single-media based retrieval, the retrieval result and user query are of the same media type. For example, you type in a text query in google, then you get many textual descriptions related to the query. In fact, users demand more diversities of the retrieval result [11]. Suppose you get lost in a strange town and you want to find the way back to your hotel. By taking a photo, cross-modal multimedia retrieval is able to return all the textual materials about where you are. Cross-modal multimedia retrieval is an exciting technology and will make our life more convenient.

Cross-modal multimedia retrieval is an exciting, yet difficult problem because multimedia objects are represented in different feature spaces. Thus the traditional single-media based methods cannot directly apply to it. The main problem is how to measure the similarity between heterogeneous objects. In this paper,

X. Lin et al. (Eds.): WISE 2013, Part I, LNCS 8180, pp. 43–56, 2013.

we address the problem by automatically learning a *heterogeneous metric* over two different spaces using labeled training data. Distance metric learning is not a new research topic. Many research efforts have been devoted to it in the last decade (e.g., [1,3,5,7,8], see [9] for a comprehensive survey). Given a training dataset of pairs of similar and dissimilar objects, distance metric learning aims to learn an optimal metric that preserves the similar/dissimilar relations among the objects. Many studies have demonstrated, both empirically and theoretically, that a learned metric can significantly improve the performance in classification, clustering and retrieval tasks [9]. However, most existing algorithms focus on learning a distance metric in a single space. Few algorithm attempt to learn a metric between different spaces.

In this paper, we propose a new approach, called Bayesian personalized ranking based heterogeneous metric learning (BPRHML). Suppose that we are learning a distance metric between spaces $\mathcal{X}$ and $\mathcal{Y}$. Let x be an object from $\mathcal{X}$, y_1 and y_2 be objects from $\mathcal{Y}$. y_1 is relevant to x, while y_2 is irrelevant to x. BPRHML computes two distances d_1 and d_2 in the transformed space. d_1 is the distance between x and y_1. d_2 is the distance between x and y_2. The key idea of BPRHML is to explicitly maximize the difference between d_1 and d_2. This will encourage the relevant objects to rank in front of the irrelevant objects. To better exploit the structure information of the heterogeneous objects, we integrate homogeneous and heterogeneous graph regularization into the objective function. Homogeneous graph regularization utilizes the similarity information inside a single space while heterogeneous graph regularization use the similarity information between different spaces. By combining them together, we can preserve smoothness of the learning result in both spaces. The objective function of BPRHML mainly consists of three terms: the loss function defined on the set of pairwise preference constraints, L2 regularization and graph regularization. We derive an efficient optimization algorithm to learn the model based on gradient descent. Experiments on the Wikipedia dataset and the corel5k image dataset show that BPRHML significantly outperforms related methods.

The rest of the paper is organized as follows: Section 2 will discusses related work. In section 3, we demonstrate preliminaries and notations. Section 4 introduces our method BPRHML. Section 5 shows the experimental results. Finally, we conclude this paper in Section 6.

2 Related Work

What lies at the core of cross-modal multimedia retrieval is to learn a metric between heterogeneous multimedia objects. In distance metric learning, a distance metric is learned from labeled training data. Typically, a linear transformation is learned to transform the data into a new space. The distance metric is then defined as the Euclidean distance in the new space. In this paper, we also adopt this definition. Most existing methods learn the metric as a Mahalanobis distance which can be represented as a positive semi-definite matrix. Given pairs of similar/dissimilar objects, approaches such as [1,3,7,8] try to learn a distance metric that keeps all the data points with the same label close, while separating data points with different label far apart. For example, in [1], the authors attempted to minimize the Mahalanobis distance between similar objects while keeping a large margin between dissimilar objects. In [5], a Mahalanobis distance is learned by maximizing the posterior probability of the training data. All of these methods focus on learning a distance metric in a single space, while in this paper a distance metric is learned between two spaces.

In heterogeneous metric learning, we usually learn two transformation matrices, which transform the heterogeneous objects into a same output space. Heterogeneous metric learning is relatively a new problem. The most related approaches to our method are [4,2,6]. Canonical correlation analysis (CCA) [10] is applied in [4] to learn a heterogeneous metric. Specially, CCA attempts to maximize the correlation between same labeled objects in the transformed space. Based on the learning results of CCA, [4] further learns a high-level semantic metric by logistic regression. Another notable approach is cross-modal factor analysis (CFA) proposed in [2]. Unlike CCA, CFA adopts a criterion of minimizing the Frobenius norm between same labeled objects in the transformed space. Both CCA and CFA consider only pairs of same labeled objects as input. They do not explicitly separate different labeled objects. To overcome this problem, Wu et.al. [6] proposed to learn two orthogonal transformation matrices by minimizing the distance between same labeled objects and maximizing the distance between different labeled objects. However, all of the above methods do not optimize for correctly ranking the retrieval results, which is important for an information retrieval system. In this paper, we attempt to preserve the partial ranking information of the training data in the transformed space.

3 Preliminaries and Notations

3.1 Cross-Modal Multimedia Retrieval

In this section, we first define the problem to be addressed. Then we introduce the notations used in this paper.

Let $\mathcal{X}$ and $\mathcal{Y}$ denote two different media types such as text, image, video. Let $\mathbb{D}_{\mathcal{X}} = \{(x_1, l_1^x), (x_2, l_2^x), \cdots, (x_m, l_m^x)\}$ and $\mathbb{D}_{\mathcal{Y}} = \{(y_1, l_1^y), (y_2, l_2^y), \cdots, (y_n, l_n^y)\}$ be two sets of multimedia objects of types $\mathcal{X}$ and $\mathcal{Y}$, respectively. l_i^x and l_i^y are labels of x_i and y_i, respectively. Our goal is to retrieve relevant x in an unlabeled dataset $\mathbb{T} = \{x_1, x_2, \cdots, x_p, y_1, y_2, \cdots, y_q\}$ in response to a query y, or vice-versa.

As for the notations, we use $\mathbf{X}$ and $\mathbf{Y}$ to denote data matrices of $\mathbb{D}_{\mathcal{X}}$ and $\mathbb{D}_{\mathcal{Y}}$. Columns of $\mathbf{X}$ and $\mathbf{Y}$ correspond to objects and rows correspond to features. x_i and y_i are the i-th column vectors of $\mathbf{X}$ and $\mathbf{Y}$, respectively. $\mathbf{U}$ and $\mathbf{V}$ are the linear transformation matrices correspond to $\mathbf{X}$ and $\mathbf{Y}$. The main notations used in this paper are summarized in Table 1.

Table 1. Notations used in this paper

Notations	Explanations
$\mathbb{D}_{\mathcal{X}}, \mathbb{D}_{\mathcal{Y}}$	training object set of types $\mathcal{X}$ and $\mathcal{Y}$
$\mathbb{T}$	test object set
l_i^x, l_i^y	labels of x_i and y_i
m, n	number of objects in $\mathbb{D}_{\mathcal{X}}$ and $\mathbb{D}_{\mathcal{Y}}$
p, q	number of objects of types $\mathcal{X}$ and $\mathcal{Y}$ in $\mathbb{T}$
α, β	regularization parameters
d^x, d^y	original dimensionality of $\mathcal{X}$ and $\mathcal{Y}$
c	dimensionality of the transformed space
$\mathbf{X}$	$d^x \times m$ data matrix of objects in $\mathbb{D}_{\mathcal{X}}$
$\mathbf{Y}$	$d^y \times n$ data matrix of objects in $\mathbb{D}_{\mathcal{Y}}$
x_i	$d^x \times 1$ column vector represents i-th object in $\mathbb{D}_{\mathcal{X}}$
y_i	$d^y \times 1$ column vector represents i-th object in $\mathbb{D}_{\mathcal{Y}}$
$\mathbf{U}$	$d^x \times c$ transformation matrix for $\mathbf{X}$
$\mathbf{V}$	$d^y \times c$ transformation matrix for $\mathbf{Y}$

3.2 Bayesian Personalized Ranking

Bayesian Personalized Ranking (BPR) [12] is a famous model of recommendation system. BPR's key idea is to use partial order of items, instead of single user-item examples to train a recommendation model. It allows the interpretation of positive-only data as partial ordering of items. When we observed a positive use-item example of user u on an item i, e.g., user u viewed or purchased item i, we assume that the user prefers this item than all other non-observed items. Formally we can extract a pairwise preference dataset $\mathcal{P}_1 : U \times I \times I$ by

$$\mathcal{P}_1 := \{(u, i, j) \mid i \in I_u^+ \wedge j \in I \setminus I_u^+\} \tag{1}$$

where U is the user set, I is the item set, I_u^+ and $I \setminus I_u^+$ are the positive item set and missing set associated with user u, respectively. Each triple $(u, i, j) \in \mathcal{P}_1$ says that user u prefers item i than j. BPR optimization criterion [12] aims to find an arbitrary model class to maximize the posterior probability over these pairs. The generic optimization criterion for Bayesian personalized ranking is :

$$\text{BPR-OPT} = -\sum_{(u,i,j) \in \mathcal{P}} \ln \sigma(\hat{x}_{uij}) + \lambda_{\Theta}(\| \Theta \|^2) \tag{2}$$

where Θ represents the parameter vector of an arbitrary model class, λ_{Θ} are model specific regularization parameters, σ is the logistic sigmoid function $\sigma(x) = 1/(1 + exp(-x))$. $\hat{x}_{uij}$ is an arbitrary real-valued function of the model parameter vector Θ which captures the special relationship between user u, item i and item j.

Note that extracting pairwise preferences constraints has been widely used in learning to rank tasks [13]. The BPR optimization criterion is actually the cross entropy cost function (logistic loss) over pairs. In fact, there are also pairwise preferences in the training dataset $\mathbb{D}_{\mathcal{X}}$ and $\mathbb{D}_{\mathcal{Y}}$. Consider the following media objects: $(x, l_1) \in \mathbb{D}_{\mathcal{X}}, (y_1, l_1) \in \mathbb{D}_{\mathcal{Y}}$ and $(y_2, l_2) \in \mathbb{D}_{\mathcal{Y}}$. Let us assume that x_1 is a query. Obviously, x_1 prefers y_1 to y_2 in the retrieval result because x_1 and y_1 are same labeled, while x_1 and y_2 are different labeled. Consequently, y_1 should rank in front of y_2 in the retrieval result.

4 BPRHML

In this section, we propose Bayesian personalized ranking based heterogeneous metric learning (BPRHML) algorithm. We first briefly review the metric learning for heterogeneous data. Then we define our objective function which consists of three terms. Finally, we derive an efficient optimization strategy. Note that BPRHML is an asymmetric model. We train different models for queries from $\mathcal{X}$ and $\mathcal{Y}$. In the rest of this paper, we assume queries are of type $\mathcal{X}$ and retrieval results are of type $\mathcal{Y}$. The algorithm for the other direction is defined analogously.

4.1 Heterogeneous Metric Learning

We can construct the set of pairwise preference constraints among the heterogeneous media objects from the training dataset $\mathbb{D}_{\mathcal{X}}$ and $\mathbb{D}_{\mathcal{Y}}$:

$$\mathcal{P}_2 = \{(x_k, y_i, y_j) \mid l_k^x = l_i^y \wedge l_k^x \neq l_j^y\} \tag{3}$$

where x_k and y_i share the same label, x_k and y_j are different labeled. Each triple (x_k, y_i, y_j) is inferred from the category labels of x_k, y_i, y_j and indicates that x_k prefers y_i to y_j in the retrieval result. Consequently, y_i should rank higher than y_j in the retrieval result. Our goal is to learn a metric between $\mathcal{X}$ and $\mathcal{Y}$ that preserves the pairwise preference constraints in $\mathcal{P}_2$.

In traditional single space metric learning, the distance metric is defined as the Mahalanobis distance between objects. The Mahalanobis distance can be viewed as a linear transformation with matrix $\mathbf{L} \in \mathbb{R}^{d^x \times c}$ followed by calculating the Euclidean distance. For an object pair (x_i, x_j), the Mahalanobis distance between them is computed as follows:

$$d(x_i, x_j) = \sqrt{(\mathbf{L}^T x_i - \mathbf{L}^T x_j)^T (\mathbf{L}^T x_i - \mathbf{L}^T x_j)} \tag{4}$$

The goal of metric learning is to learn the linear transformation matrix $\mathbf{L}$ from the set of similar/dissimilar constraints. However, in heterogeneous metric learning, objects x_i and y_j are coming from two heterogeneous spaces $\mathcal{X}$ and $\mathcal{Y}$ with different features (e.g., dimensions). The similarity relation between heterogeneous data is not a metric, hence, does not fall into the standard framework of metric learning. It is not trivial to define a metric between two heterogeneous spaces. Our proposal is to learn two linear transformations, which transform heterogeneous objects into a same output space. More specially, let $\mathbf{U} \in \mathbb{R}^{d^x \times c}$ be the transformation matrix for $\mathbf{X} \in \mathbb{R}^{d^x \times m}$ and $\mathbf{V} \in \mathbb{R}^{d^y \times c}$ be the transformation matrix for $\mathbf{Y} \in \mathbb{R}^{d^y \times n}$, d^x is the original dimensionality of $\mathcal{X}$ and d^y is the original dimensionality of $\mathcal{Y}$, m and n are the number of media objects in $\mathbb{D}_{\mathcal{X}}$ and $\mathbb{D}_{\mathcal{Y}}$, c is the dimensionality of the transformed space. For an object pair (x_i, y_j), we define the heterogeneous distance as the Euclidean distance in the transformed space:

$$d(x_i, y_j) = \sqrt{(\mathbf{U}^T x_i - \mathbf{V}^T y_j)^T (\mathbf{U}^T x_i - \mathbf{V}^T y_j)} \tag{5}$$

Our formulation naturally extends conventional distance metric learning from one single space to two different spaces. Single space metric learning can be viewed as a special case of our formulation in which $\mathbf{U} = \mathbf{V}$. We aim to learn the parameter matrices $\mathbf{U}$ and $\mathbf{V}$ that preserve the pairwise preference constraints in $\mathcal{P}_2$.

4.2 Objective Function

We construct an objective function which consists of three terms for heterogeneous metric learning as follow:

$$\underset{\mathbf{U},\mathbf{V}}{\operatorname{argmin}} \; l(\mathbf{U}, \mathbf{V}) + \alpha s(\mathbf{U}, \mathbf{V}) + \beta g(\mathbf{U}, \mathbf{V}) \tag{6}$$

where $l(\mathbf{U}, \mathbf{V})$ is the loss function defined on the set of pairwise preference constraints, $s(\mathbf{U}, \mathbf{V})$ is the L2 regularization, $g(\mathbf{U}, \mathbf{V})$ is the graph regularization, α and β are regularization parameters.

Loss Function. We argue that the loss function should optimize for correctly ranking the retrieval result. Minimizing the loss function will encourage relevant objects to rank in front of irrelevant objects, i.e, preserving the pairwise preference constraints in $\mathcal{P}_2$. Our proposal is to take the advantage of the Bayesian personalized ranking model introduced in section 3.2. One simple formulation of the loss function is as follows:

$$l(\mathbf{U}, \mathbf{V}) = -\tfrac{1}{2} \sum_{k=1}^{m} \sum_{i \in \mathbf{Y}_k^+} \sum_{j \in \mathbf{Y}_k^-} \ln \sigma(\hat{x}_{kij}) \tag{7}$$

where x_k is an object from $\mathbf{X}$, $\mathbf{Y}_k^+$ is the set of objects in $\mathbf{Y}$ that are same labeled with x_k, $\mathbf{Y}_k^-$ is the set of objects in $\mathbf{Y}$ that are different labeled with x_k, and $\hat{x}_{kij}$ is defined as follows:

$$\hat{x}_{kij} = \parallel \mathbf{U}^T x_k - \mathbf{V}^T y_i \parallel^2 - \parallel \mathbf{U}^T x_k - \mathbf{V}^T y_j \parallel^2 \tag{8}$$

where $\parallel \boldsymbol{.} \parallel^2$ denotes the square of L2 norm. Intuitively, for a given triple (k, i, j), minimizing $l(\mathbf{U}, \mathbf{V})$ will result in maximizing $\hat{x}_{kij}$, i.e., encouraging relevant object to rank in front of irrelevant object. However, for a training dataset which consists of m objects of type $\mathcal{X}$ and n objects of type $\mathcal{Y}$, there are possibly $O(m \times n^2)$ pairwise preference constraints in $\mathcal{P}_2$. In case of large m and n, the huge number of pairwise preference constraints in $\mathcal{P}_2$ will affect the efficiency of the optimization algorithm.

To handle the above issue, we propose to construct two *representative object* sets out of the relevant and irrelevant object sets, respectively. More specially, let

$x_k \in \mathbb{D}_{\mathcal{X}}$ be a query, $\mathbf{Y}_k^+ \subseteq \mathbb{D}_{\mathcal{Y}}$ and $\mathbf{Y}_k^- \subseteq \mathbb{D}_{\mathcal{Y}}$ be the corresponding relevant and irrelevant object sets. We construct two representative object set D_k^+ and D_k^- out of $\mathbf{Y}_k^+$ and $\mathbf{Y}_k^-$, respectively. Intuitively, objects in D_k^+ are representatives of objects in $\mathbf{Y}_k^+$ and objects in D_k^- are representatives of objects in $\mathbf{Y}_k^-$. Note that objects in $\mathbf{Y}_k^+$ share the same label, while objects in $\mathbf{Y}_k^-$ are different labeled. To construct D_k^+, we cluster the objects in $\mathbf{Y}_k^+$. Suppose that we have built M clusters $(C_1^+, \cdots, C_M^+)$ by applying a clustering algorithm such as K-means. Then we define D_k^+ to be the centroid of each cluster:

$$D_k^+ = \{cen(C_i^+) \mid 1 \le i \le M\} \tag{9}$$

$$cen(C_i^+) = \frac{1}{|C_i^+|} \sum_{y \in C_i^+} y \tag{10}$$

where $|\ \boldsymbol{.}\ |$ denotes set cardinality. As for D_k^-, we first divide $\mathbf{Y}_k^-$ into N clusters according to the labels, i.e., same labeled objects form a cluster. N is the number of labels in $\mathbf{Y}_k^-$. Assume that we have built N clusters $(C_1^-, \cdots, C_N^-)$. Then we also define D_k^- to be the centroid of each cluster:

$$D_k^- = \{cen(C_i^-) \mid 1 \le i \le N\} \tag{11}$$

$$cen(C_i^-) = \frac{1}{|C_i^-|} \sum_{y \in C_i^-} y \tag{12}$$

In fact, we can further perform clustering on D_k^- to reduce the number of representative objects in D_k^-. Note that we are not the first to define representative object as the centroid of corresponding object set. In [17], the authors attempted to learn latent factors by maximizing the marginal utility between user choice and the *average* of non-choices. With a slight abuse of notation, we will also denote an object from D_k^+ or D_k^- by y_i or y_j. Based on D_k^+ and D_k^-, we can construct the new set of pairwise preference constraints as follows:

$$\mathcal{P}_2' = \{(x_k, y_i, y_j) \mid 1 \le k \le m \wedge i \in D_k^+ \wedge j \in D_k^-\} \tag{13}$$

By constructing the representative objects, we reduce the number of pairwise preference constraints to $O(m \times M \times N)$, where $M \ll n$ and $N \ll n$. The loss function is defined to preserve the constraints in $\mathcal{P}_2'$:

$$l(\mathbf{U}, \mathbf{V}) = -\frac{1}{2} \sum_{k=1}^{m} \sum_{i \in D_k^+} \sum_{j \in D_k^-} \ln \sigma(\hat{x}_{kij}) \tag{14}$$

where $\hat{x}_{kij}$ is defined similar to (8). Intuitively, $\sigma(\hat{x}_{kij})$ defines the probability for y_i to rank in front of y_j in the retrieval result.

L2 Regularization. We define the L2 regularization as follows:

$$s(\mathbf{U}, \mathbf{V}) = \frac{1}{2}(\parallel \mathbf{U} \parallel_F^2 + \parallel \mathbf{V} \parallel_F^2) \tag{15}$$

$\parallel \mathbf{U} \parallel_F^2$ and $\parallel \mathbf{V} \parallel_F^2$ are the square of Frobenius norm of $\mathbf{U}$ and $\mathbf{V}$, respectively. L2 regularization is widely used to reduce overfitting.

Graph Regularization. Graph regularization has been widely used in dimensionality reduction [19], clustering [20] and semi-supervised learning [21]. The key assumption of graph regularization is that if two media objects are similar, they should also be close to each other in the transformed space. In heterogeneous metric learning, we have similarity constraints in single modality and across modalities. Therefore, we intend to define *homogeneous* graph regularization and *heterogeneous* graph regularization, respectively. Homogeneous graph regularization captures the similarity information inside a single modality, i.e, $\mathbf{X}$ or $\mathbf{Y}$. In the following, we define the homogeneous graph regularization for $\mathbf{X}$. The homogeneous graph regularization for $\mathbf{Y}$ is defined analogously. In homogeneous space, we can use both label information and distance information. Following [18], we define the homogeneous neighbourhood of an object x_i, denoted as $\mathcal{N}_i$, to be the k nearest neighbours, determined by Euclidean distance, that share the same label with x_i. We treat the same labeled objects outside $\mathcal{N}_i$ as outliers and ignore them. We define an undirected and symmetric data graph $G_x = (V_x, \mathbf{W}_x)$ on $\mathbf{X}$. V_x is the set of objects in $\mathbf{X}$. $\mathbf{W}_x$ is a $m \times m$ matrix and each element w_{ij} of $\mathbf{W}_x$ denotes the similarity information between the i-th media object and j-th media object of $\mathbf{X}$. Based on the homogeneous neighbourhood, $\mathbf{W}_x$ is defined as follows:

$$w_{ij} = \begin{cases} 1, \ (x_j \in \mathcal{N}_i \vee x_i \in \mathcal{N}_j) \wedge i \neq j \\ 0, \ otherwise \end{cases} \tag{16}$$

w_{ii} is set to 0 to avoid self-reinforcement. Let $\mathbf{T} = \mathbf{U}^T\mathbf{X}$ and t_i be the i-th column of $\mathbf{T}$. Intuitively, $\mathbf{T}$ represents all media objects of $\mathbf{X}$ in the transformed space. The homogeneous graph regularization is defined as follows:

$$\begin{aligned} \mathcal{O}_1 &= \tfrac{1}{4} \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} \parallel \tfrac{t_i}{\sqrt{d_{ii}}} - \tfrac{t_j}{\sqrt{d_{jj}}} \parallel^2 \\ &= \tfrac{1}{2} tr(\mathbf{T}\mathbf{L}_x\mathbf{T}^T) \end{aligned} \tag{17}$$

where $\mathbf{L}_x = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}_x\mathbf{D}^{-1/2}$ is called the graph Laplacian with $\mathbf{D}$ being a diagonal matrix whose diagonal are row sums of $\mathbf{W}_x$, $d_{ii} = \sum_j w_{ij}$, $\mathbf{I}$ is an $m \times m$ identity matrix, and $tr(.)$ denotes the trace of a matrix.

Unlike in homogeneous space, we have only label information in heterogeneous spaces. So we construct the data graph $G_{xy} = (V_{xy}, \mathbf{W}_{xy})$ between $\mathbf{X}$ and $\mathbf{Y}$ from the labels. $\mathbf{W}_{xy}$ is a $m \times n$ matrix and each element w_{ij} of $\mathbf{W}_{xy}$ denotes the similarity information between the i-th media object of $\mathbf{X}$ and the j-th media object of $\mathbf{Y}$. $\mathbf{W}_{xy}$ is defined as follows:

$$w_{ij} = \begin{cases} 1, \ l_x^i = l_y^j \wedge 1 \leq i \leq m \wedge 1 \leq j \leq n \\ 0, \ otherwise \end{cases} \tag{18}$$

Let $\mathbf{S} = \mathbf{V}^T\mathbf{Y}$ and s_i be the i-th column of $\mathbf{S}$. $\mathbf{S}$ represents all media objects of $\mathbf{Y}$ in the transformed space. The heterogeneous graph regularization is defined as follows:

$$\begin{aligned} \mathcal{O}_2 &= \tfrac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} \parallel \tfrac{t_i}{\sqrt{d_{ii}^x}} - \tfrac{s_j}{\sqrt{d_{jj}^y}} \parallel^2 \\ &= \tfrac{1}{2} tr(\mathbf{T}\mathbf{T}^T) + \tfrac{1}{2} tr(\mathbf{S}\mathbf{S}^T) - tr(\mathbf{T}\mathbf{D}_x^{-1/2}\mathbf{W}_{xy}\mathbf{D}_y^{-1/2}\mathbf{S}^T) \end{aligned} \tag{19}$$

where $\mathbf{D}_x$ is a $m \times m$ diagonal matrix whose diagonal are row sums of $\mathbf{W}_{xy}$, $d^x_{ii} = \sum_j w_{ij}$, $\mathbf{D}_y$ is a $n \times n$ diagonal matrix whose diagonal are column sums of $\mathbf{W}_{xy}$, $d^y_{ii} = \sum_j w_{ji}$. In summary, the graph regularization $g(\mathbf{U}, \mathbf{V})$ is defined to be:

$$\begin{aligned} g(\mathbf{U},\mathbf{V}) &= \tfrac{1}{2}tr(\mathbf{T}\mathbf{L}_x\mathbf{T}^T) + \tfrac{1}{2}tr(\mathbf{S}\mathbf{L}_y\mathbf{S}^T) + \tfrac{1}{2}tr(\mathbf{T}\mathbf{T}^T) \\ &\quad + \tfrac{1}{2}tr(\mathbf{S}\mathbf{S}^T) - tr(\mathbf{T}\mathbf{D}_x^{-1/2}\mathbf{W}_{xy}\mathbf{D}_y^{-1/2}\mathbf{S}^T) \\ &= \tfrac{1}{2}tr(\mathbf{U}^T\mathbf{X}\mathbf{L}'_x\mathbf{X}^T\mathbf{U}) + \tfrac{1}{2}tr(\mathbf{V}^T\mathbf{Y}\mathbf{L}'_y\mathbf{Y}^T\mathbf{V}) \\ &\quad - tr(\mathbf{U}^T\mathbf{X}\mathbf{D}_x^{-1/2}\mathbf{W}_{xy}\mathbf{D}_y^{-1/2}\mathbf{Y}^T\mathbf{V}) \end{aligned} \tag{20}$$

where $\mathbf{L}'_x = \mathbf{L}_x + \mathbf{I}$, $\mathbf{L}'_y = \mathbf{L}_y + \mathbf{I}$, $\mathbf{L}_y$ is the graph Laplacian corresponding to $\mathbf{Y}$. Minimizing $g(\mathbf{U}, \mathbf{V})$ will encourage the smoothness of the transformation over both modalities.

4.3 Optimization Strategy

Firstly, we show how to initialize $\mathbf{U}$ and $\mathbf{V}$. Among all the related methods introduced in section 2, CFA shares the same assumption with our method. CFA also assumes there are two linear transformations that transform the heterogeneous objects into a same output space. Then the heterogeneous distance metric is defined as the Euclidean distance in the transformed space. Consequently, we propose to initialize $\mathbf{U}$ and $\mathbf{V}$ with the learning result of CFA. More specially, CFA optimizes the following objective function:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \parallel \mathbf{U}^T\mathbf{X_1} - \mathbf{V}^T\mathbf{Y_1} \parallel_F^2 \\ &s.t. \quad \mathbf{U}\mathbf{U}^T = \mathbf{I}, \quad \mathbf{V}\mathbf{V}^T = \mathbf{I} \end{aligned} \tag{21}$$

where $\mathbf{X_1}$ and $\mathbf{Y_1}$ are two object matrices, which consist of row-by-row coupled samples of two media types, and $\mathbf{I}$ is identity matrix of corresponding size. We can rewrite the above objective function as follows:

$$\parallel \mathbf{U}^T\mathbf{X_1} - \mathbf{V}^T\mathbf{Y_1} \parallel^2 = tr(\mathbf{X_1}^T\mathbf{X_1}) + tr(\mathbf{Y_1}^T\mathbf{Y_1}) - 2tr(\mathbf{X_1}^T\mathbf{U}\mathbf{V}^T\mathbf{Y_1}) \tag{22}$$

We can easily see from the above that matrices $\mathbf{U}$ and $\mathbf{V}$ that maximize $tr(\mathbf{X_1}^T\mathbf{U}\mathbf{V}^T\mathbf{Y_1})$ will minimize (20). It can be shown [16] that such matrices are given by singular value decomposition:

$$\mathbf{X_1}\mathbf{Y_1}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{23}$$

Instead of adopting stochastic gradient descent like in Bayesian personalized ranking [12], we derive our optimization algorithm based on gradient descent. There are two reasons for this choice: 1) by introducing representative object, the number of pairwise preference constraints reduces to $O(m \times M \times N)$. 2) one update on a preference triple (k, i, j) will affect all variables, i.e., $\mathbf{U}$ and $\mathbf{V}$, in the objective function. Let $\mathcal{F}(\mathbf{U}, \mathbf{V})$ denote the objective function in (6). Once the initial value of $\mathbf{U}$ and $\mathbf{V}$ are computed, we update $\mathbf{U}$ and $\mathbf{V}$ in each iteration by the following gradients:

Algorithm 1. Learning procedure of BPRHML

```
Input: training data matrices $\mathbf{X}$ and $\mathbf{Y}$, learning rate $\eta$, regularization parameters $\alpha$ and $\beta$, dimensionality $K$ of the transformed space
Output: transformation matrices $\mathbf{U}$ and $\mathbf{V}$
Construct the data graph matrices $\mathbf{W}_x$, $\mathbf{W}_y$ and $\mathbf{W}_{xy}$
Compute the Laplacian matrices $\mathbf{L}_x$ and $\mathbf{L}_y$ based on $\mathbf{W}_x$ and $\mathbf{W}_y$
Compute the diagonal matrices $\mathbf{D}_x$ and $\mathbf{D}_y$ based on $\mathbf{W}_{xy}$
Compute the temporary matrices $\mathbf{U}'$ and $\mathbf{V}'$ by $\mathbf{X}\mathbf{Y}^T = \mathbf{U}'\mathbf{\Sigma}\mathbf{V}'^T$
Initialize $\mathbf{U}$ and $\mathbf{V}$ with the first $K$ columns of $\mathbf{U}'$ and $\mathbf{V}'$, respectively
repeat
    update $\mathbf{U}$ by $\mathbf{U} \leftarrow \mathbf{U} - \eta\frac{\partial\mathcal{F}}{\partial\mathbf{U}}$
    update $\mathbf{V}$ by $\mathbf{V} \leftarrow \mathbf{V} - \eta\frac{\partial\mathcal{F}}{\partial\mathbf{V}}$
until convergence
```

$$\begin{aligned}\frac{\partial\mathcal{F}}{\partial\mathbf{U}} &= \sum_{k=1}^{m}\sum_{i\in D_k^+}\sum_{j\in D_k^-}\frac{1}{1+exp(\hat{x}_{kij})}x_k(y_i^T - y_j^T)\mathbf{V} + \alpha\mathbf{U} \\ &+\beta\mathbf{X}\mathbf{L}'_x\mathbf{X}^T\mathbf{U} - \beta\mathbf{X}\mathbf{D}_x^{-1/2}\mathbf{W}_{xy}\mathbf{D}_y^{-1/2}\mathbf{Y}^T\mathbf{V} \qquad (24)\\ \frac{\partial\mathcal{F}}{\partial\mathbf{V}} &= \sum_{k=1}^{m}\sum_{i\in D_k^+}\sum_{j\in D_k^-}\frac{1}{1+exp(\hat{x}_{kij})}((y_i - y_j)x_k^T\mathbf{U} + (y_jy_j^T - y_iy_i^T)\mathbf{V}) \\ &+\alpha\mathbf{V} + \beta\mathbf{Y}\mathbf{L}'_y\mathbf{Y}^T\mathbf{V} - \beta\mathbf{Y}\mathbf{D}_y^{-1/2}\mathbf{W}_{xy}^T\mathbf{D}_x^{-1/2}\mathbf{X}^T\mathbf{U} \qquad (25)\end{aligned}$$

We use a constant learning rate to update the transformation matrices. The process of estimating $\mathbf{U}$ and $\mathbf{V}$ is described in algorithm 1.

5 Experiments

We conduct experiments on two publicly available dataset to compare the performance of our method with other state-of-the-art methods.

5.1 Datasets and Evaluation Criteria

Cross-modal multimedia retrieval is relatively a new problem. There are few publicly available benchmark datasets. To the best of our knowledge, the Wikipedia dataset proposed by Rasiwasia et.al. [4] is the only publicly available dataset specially collected for cross-modal multimedia retrieval. To further evaluate the performance of our method, we also construct experiment on the corel5k image dataset, which is widely used in image annotation [14,15]. In the following, we will introduce the above two dataset in detail.

Wikipedia dataset[1] contains documents that are selected sections from the Wikipedia's featured articles collection. This is a continually updated collection of 2700 articles that have been selected and reviewed by Wikipedia's editors since 2009. The article generally have multiple sections and pictures. Each article

[1] http://www.svcl.ucsd.edu/projects/crossmodal/

is split into sections based on section headings, and assign each image to the section in which it was placed by the authors. The final dataset contains a total of 2866 documents, which are text-image pairs, annotated with a label from the vocabulary of 10 semantic categories. The dataset is randomly split into a training set of 2173 documents and a test set of 693 documents.

Corel5k dataset[2] is a widely used benchmark for image annotation. The dataset consists of 4500 training images and 500 test images and there are 260 possible keywords. Each image is annotated with 1-5 key words. The average keywords per image is 3.4. We treat each keyword as a *pseudo document*. The pseudo document is relevant to an image if the corresponding keyword is used to annotate the image. The pseudo document is represented in a space in which each dimension corresponds to a keyword and the dimensionality is 260. We use co-occurrence of keywords in the annotation matrix as feature vectors. Image representation is based on the popular scale invariant feature transformation (SIFT). We perform principal component analysis on both the pseudo document matrix and the image matrix and preserve 85% variance. The final dimensionality of pseudo document matrix and image matrix is 18 and 100, respectively.

Similar to [4] and [6], we adopt Mean Average Precision (MAP) as the evaluation criteria. The MAP score is the average precision at the ranks where recall changes. It is widely used in the information retrieval literature.

5.2 Comparison Settings

In order to show the effectiveness of our approach, we compare the results with the following five baseline methods.

1. **Random:** Randomly retrieving the results.
2. **CCA:** Canonical correlation analysis is used in [4] to learn two transformation matrices that maximize the correlation between two sets of heterogeneous objects.
3. **CFA:** CFA learns two linear transformation matrices [2]. Unlike CCA, CFA optimizes for minimizing the Frobenius norm between pairwise objects in the transformed space.
4. **SCM:** SCM is proposed by Rasiwasia et.al. [4]. CCA is first applied to learn two maximally correlated subspaces. Then it applies logistic regression to learn a high-level semantic representation of the media objects.
5. **MSmethod:** MSmethod is currently state-of-the-art method [6]. It learns two orthogonal transformation matrices by minimizing the distance between relevant objects and maximizing the distance between irrelevant objects.

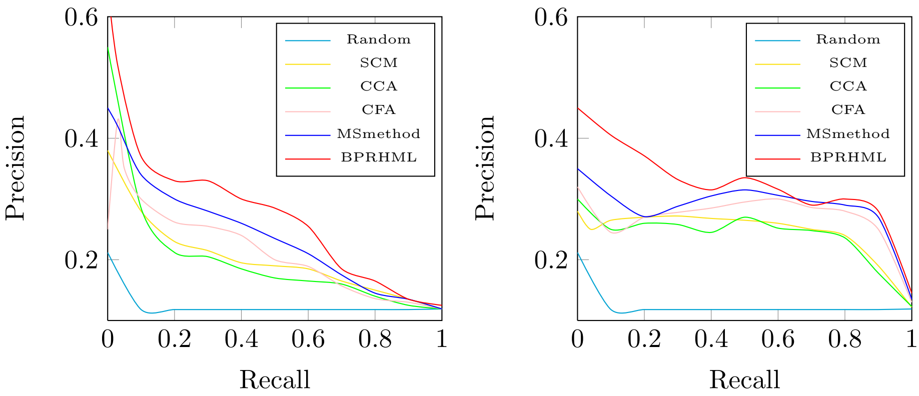
5.3 Experimental Results

In this section, we compare the performance of our method with the above five baseline methods on the Wikipedia dataset and the corel5k dataset.

[2] `http://lear.inrialpes.fr/people/guillaumin/data.php`

Table 2. MAP values on Wikipedia and Corel5k dataset

Task	Random	CCA	CFA	SCM	MSmethod	BPRHML
Image→Text	0.118	0.249	0.279	0.277	0.282	**0.299**
Text→Image	0.118	0.196	0.231	0.226	0.238	**0.265**
Image→Keyword	0.054	0.117	0.112	0.126	0.135	**0.179**
Keyword→Image	0.061	0.125	0.136	0.131	0.147	**0.167**
Average	0.088	0.172	0.189	0.190	0.201	**0.228**



(a) Text Query (b) Image Query

Fig.1. Precision recall curves on Wikipedia dataset

Table 2 shows the MAP values of our method and all other baseline methods. The upper part of Table 2 shows the MAP values on the Wikipedia dataset. The low part of Table 2 shows the MAP values on the corel5k dataset. The better results are shown in bold. To make a fair comparison, we tune all methods to their best according to the 5-fold cross validation on the training dataset. As for the parameters of BPRHML, we set $\alpha = 86$, $\beta = 7.1$ for the task of retrieving text by image query and set $\alpha = 1000$, $\beta = 0.001$ for the task of retrieving image by text query. For both tasks, we set $k = 5$ for K-means clustering and set $k = 50$ for computing the homogeneous neighbourhood.

Due to the factor that BPRHML explicitly optimizes for correctly ranking the retrieval result, it outperforms the compared baselines on both datasets and tasks consistently. We observe that the performance of most methods decrease on the corel5k dataset. It is reasonable since there are only keywords, rather than text in the corel5k dataset. So the corel5k dataset contains less textual information than the Wikipedia dataset, which makes it more challenge. However, BPRHML still significantly outperforms other baseline methods on this challenge dataset. It can be seen from Table 2 that the performance of MSmethod is comparable to our method. It is not surprise because MSmethod considers both similar and dissimilar information while the other baseline methods consider only similar information. CFA outperforms CCA and SMN in most of the case, this demonstrates that the definition of heterogeneous distance metric as the Euclidean distance after two linear transformations is effective for cross-modal multimedia retrieval. In addition, we observe that SCM always outperforms CCA. This suggests that we can further improve the performance of our method by learning a high-level semantic representation [4] of the heterogeneous objects. We leave it as future work. Further analysis of the results is presented in Figure 1, which shows the PR curve of all approaches for both image and text queries on the Wikipedia dataset. We can see from Figure 1 that BPRHML achieves high precision at most levels of recall.

6 Conclusion

In this paper, we propose a Bayesian personalized ranking based heterogeneous metric learning (BPRHML) algorithm to learn the distance metric between heterogeneous objects. We assume that there are two linear transformation matrices which transform the heterogeneous objects into a same output space. The heterogeneous distance metric is then defined as the Euclidean distance in the transformed space. We argue that good objective function should optimize for correctly ranking the retrieval result. So we formulate an objective function which can preserve the pairwise preference constraints in the training data. To further exploiting the structure information contained in the training data, we integrate homogeneous and heterogeneous graph regularization into the objective function. Experiments on benchmark datasets demonstrate the effectiveness of our method. The experimental datasets of this paper contain only text and images. In the future, we intend to evaluate our method on more multimedia types, such as audio and video. We will also learn a high-level semantic representation of the heterogeneous objects based on the learning result of BPRHML.

Acknowledgments. We would like to thank all anonymous reviewers for their helpful comments. This work is supported in part by NSFC grant 60970045 and China National 973 project 2013CB329305.

References

1. Xing, E., Ng, A., Jordan, M., Russell, S.: Distance metric learning with application to clustering with side-information. In: NIPS 2002, pp. 505–512 (2002)
2. Li, D., Dimitrova, N., Li, M., Sethi, I.K.: Multimedia content processing through cross-modal association. In: Proceedings of the Eleventh ACM International Conference on Multimedia, pp. 604–611 (2003)
3. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbour classification. In: NIPS 2006, pp. 1475–1482 (2006)
4. Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A New Approach to Cross-Modal Multimedia Retrieval. In: Proceedings of the Eighteenth International Conference on Multimedia, pp. 251–260 (2010)

5. Liu, Y., Rong, J., Rahul, S.: Bayesian Active Distance Metric Learning. In: Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI 2007), pp. 442–449 (2007)
6. Wu, W., Xu, J., Li, H.: Learning Similarity Function between Objects in Heterogeneous Spaces. Microsoft Research Technique Report (2010)
7. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.: Information-Theoretic Metric Learning. In: ICML 2007, pp. 209–216 (2007)
8. Hoi, S.C.H., Liu, W., Chang, F.: Semi-Supervised Distance Metric Learning for Collaborative Image Retrieval. In: CVPR 2008, pp. 1–7 (2008)
9. Liu, Y.: Distance Metric Learning: A Comprehensive Survey, School of Computer Science, Carnegie Mellon University (2006)
10. Timm, N.: Applied multivariate analysis. Springer (2002)
11. Liu, J., Xu, C.S., Lu, H.Q.: Cross-media retrieval: state-of-the-art and open issues. International Journal of Multimedia Intelligence and Security 1(1), 33–52 (2010)
12. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: Bayesian Personalized Ranking from Implicit Feedback. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 452–461 (2009)
13. Liu, T.: Learning to rank for information retrieval. Foundations and Trends in Information Retrieval 3(3), 225–331 (2009)
14. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV 2009, pp. 309–316 (2009)
15. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: CVPR 2010, pp. 902–909 (2010)
16. Krzanowski, W.: Principles of multivariate analysis. Oxford University Press, Oxford (1988)
17. Yang, S.H., Long, B., Smola, A., Zha, H.Y., Zheng, Z.H.: Collaborative Competitive Filtering: Learning Recommender Using Context of User Choice. In: SIGIR 2011, pp. 295–304 (2011)
18. Wang, F., Sun, J., Li, T., Anerousis, N.: Two heads better than one: Metric+active learning and its applications for it service classification. In: ICDM 2009, pp. 1022–1027 (2009)
19. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computing, 1373–1396 (2003)
20. Cai, D., He, X., Han, J., Huang, T.: Graph regularized non-negative matrix factorization for data representation. IEEE Transaction on Pattern Analysis and Machine Intelligence (2010)
21. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled an unlabeled examples. The Journal of Machine Learning Research 7, 2399–2434 (2006)