Semi-Supervised Distance Metric Learning for Collaborative Image Retrieval

Steven C.H. Hoi
School of Computer Engineering
Nanyang Technological University
chhoi@ntu.edu.sg

Wei Liu and Shih-Fu Chang
Department of Electrical Engineering
Columbia University
{wliu,sfchang}@ee.columbia.edu

Abstract

Typical content-based image retrieval (CBIR) solutions with regular Euclidean metric usually cannot achieve satisfactory performance due to the semantic gap challenge. Hence, relevance feedback has been adopted as a promising approach to improve the search performance. In this paper, we propose a novel idea of learning with historical relevance feedback log data, and adopt a new paradigm called "Collaborative Image Retrieval" (CIR). To effectively explore the log data, we propose a novel semi-supervised distance metric learning technique, called "Laplacian Regularized Metric Learning" (LRML), for learning robust distance metrics for CIR. Different from previous methods, the proposed LRML method integrates both log data and unlabeled data information through an effective graph regularization framework. We show that reliable metrics can be learned from real log data even they may be noisy and limited at the beginning stage of a CIR system. We conducted extensive evaluation to compare the proposed method with a large number of competing methods, including 2 standard metrics, 3 unsupervised metrics, and 4 supervised metrics with side information.

1. Introduction

Determination of appropriate distance metrics plays a key role in building an effective content-based image retrieval (CBIR) system. Regular CBIR systems usually adopt Euclidean metrics for computing distances between images that are represented in some vector space. Unfortunately, Euclidean distance is often inadequate primarily because of the well-known semantic gap between low-level features and high-level semantic concepts [18].

In response to the semantic gap challenge, relevance feedback techniques have been extensively studied in CBIR [9, 10] and are shown effective in some applications. However, they also suffer from some drawbacks. The most obvious one is the addition of communication overhead imposed on the systems and users. CBIR systems with relevance feedback often require a non-trivial number of iterations before improved search results are obtained. This makes the process inefficient and unattractive for online applications.

Beyond relevance feedback, several promising directions emerge in addressing the semantic gap issue. For example, image annotation attempts to infer semantic concepts from low-level image contents. Recent works [4] have shown interesting progress, though major challenges still remain. In this work, we consider an alternative paradigm, called "Collaborative Image Retrieval" (CIR), for attacking the semantic gap challenge. CIR has attracted growing interest recently [13, 17]. It avoids the aforementioned major overhead on users in image retrieval tasks, by leveraging the historical log data of user relevance feedback collected from real CBIR systems over a long period of time. The relevance feedback information is not limited to only the data collected from the current session of image search. Instead, all of the historical data from prior interaction by a large group of users is utilized to discover useful information.

The key for CIR is to find an effective way of utilizing the log data of user relevance feedback so that the semantic gap can be effectively bridged. In this paper, we explore the log data for learning distance metrics required in image retrieval tasks. Recently, learning distance metrics from log data (or called "side information" [23]) has been actively studied in machine learning. In this paper, we propose novel formulation and develop effective algorithms for distance metrics learning (DML) in the context of CIR.

Regular DML techniques are sensitive to noise and unable to learn a reliable metric when only a small amount of log data is available. In this paper, we propose a new semi-supervised distance metric learning scheme for incorporating unlabeled data in the distance metric learning task. Specifically, we develop a novel Laplacian Regularized Metric Learning (LRML) algorithm, to integrate the unlabeled data information through a graph regularization learning framework. The LRML algorithm is formulated as a Semidefinite Program (SDP), which can be solved to find global optimum efficiently by existing convex opti-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

mization techniques. Here we highlight the major contributions in this paper: (1) a novel regularization framework for distance metric learning and a new semi-supervised metric learning algorithm; (2) a comprehensive study of a new CIR paradigm with the exploration of real log data; (3) an extensive evaluation of a number of competing metric learning methods for CIR applications.

The rest of this paper is organized as follows. Section 2 includes review of related work. Section 3 defines the distance metric learning problem and formulates the proposed semi-supervised distance metric learning technique for CIR applications. Section 4 presents our experimental evaluations on some testbed with real user log data collected from a CBIR system. Section 5 concludes this paper.

2. Related Work

Our work is mainly related to two groups of research. One is the work for exploring the log data of user relevance feedback in CBIR. The other is the distance metric learning research in machine learning. We briefly review some representative work in both sides.

In recent years, there are some emerging research interests for exploring the historical log data of user relevance feedback in CBIR. Hoi et al. [13] proposed the log-based relevance feedback with support vector machines (SVM) techniques by engaging the user feedback log data in traditional online relevance feedback tasks. Similarly, there were some other efforts in exploring the log data with other machine learning techniques, such as the manifold learning solution [8] and the coupled SVMs [12]. Different from previous work, we study distance metric learning for exploring user log data that avoids the needs of using online relevance feedback explicitly.

The other major group of related work is the distance metric learning research in machine learning, which can be classified into three major categories. One is unsupervised learning approaches, most of which attempt to find low-dimensional embeddings from high-dimensional input data. Some well-known techniques include classical Principal Component Analysis (PCA) [5] and Multidimensional Scaling (MDS) [3]. Some manifold based approaches study nonlinear techniques, such as Locally Linear Embedding (LLE) [16] and Isomap [20], etc. Another category is supervised metric learning techniques for classification. These methods usually learn metrics from training data associated with explicit class labels. The representative techniques include Fisher Linear Discriminant Analysis (LDA) [5] and some recently proposed methods, such as Neighbourhood Components Analysis (NCA) [14], Maximally Collapsing Metric Learning [7], metric learning for Large Margin Nearest Neighbor classification (LMNN) [22], and Local Distance Metric Learning [24], etc.

Our DML work is closer to the third category of DML, which learns distance metrics from the log data of pairwise constraints, or known as "side information" [23], in which each constraint indicates if two data points are relevant (similar) or irrelevant (dissimilar) in a particular learning task. A well-known DML approach was proposed by Xing et al. [23], who formulated the task as a convex optimization problem, and applied the solution to clustering tasks. Following their work, there are a group of emerging DML techniques proposed in this direction. For example, Relevance Component Analysis (RCA) learns a global linear transformation by exploiting only the equivalent constraints [1]. Discriminant Component Analysis (DCA) improves the RCA by incorporating the negative constraints [11]. Recently, Si et al. proposed a regularized metric learning method for CIR applications [17]. In this paper, we propose a new semi-supervised distance metric learning framework for learning effective and reliable metrics by incorporating the unlabeled data in DML.

3. Semi-Supervised Distance Metric Learning

3.1. Problem Definition

Assume that we are given a set of n data points $\mathcal{C} = \{\mathbf{x}_i\}_{i=1}^n \subseteq \mathbb{R}^m$, and two sets of pairwise constraints among the data points:

$$\begin{aligned}\mathcal{S} &= \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are judged to be relevant}\}\\ \mathcal{D} &= \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are judged to be irrelevant}\}\end{aligned}$$

where $\mathcal{S}$ is the set of similar pairwise constraints, and $\mathcal{D}$ is the set of dissimilar pairwise constraints. Each pairwise constraint $(\mathbf{x}_i, \mathbf{x}_j)$ indicates if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are relevant or irrelevant judged by users under some application context.

For any two given data points $\mathbf{x}_i$ and $\mathbf{x}_j$, let $d(\mathbf{x}_i, \mathbf{x}_j)$ denote the distance between them. To compute the distance, let $\mathbf{A} \in \mathbb{R}^{m\times m}$ be the distance metric, we can then express the formula of distance measure as follows:

$$\begin{aligned}d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) &= \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{A} (\mathbf{x}_i - \mathbf{x}_j)}\\ &= \sqrt{\mathbf{tr}(\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top)}\,, \qquad (1)\end{aligned}$$

where $\mathbf{A}$ is a symmetric matrix of size $m \times m$, and **tr** stands for the *trace* operator. In general, $\mathbf{A}$ is a valid metric if and only if it satisfies the non-negativity and the triangle inequality properties. In other words, the matrix $\mathbf{A}$ must be positive semi-definite, i.e., $\mathbf{A} \succeq 0$. Generally, the matrix $\mathbf{A}$ parameterizes a family of Mahalanobis distances on the vector space $\mathbb{R}^m$. Specifically, when setting $\mathbf{A}$ to be an identity matrix $\mathbf{I}_{m\times m}$, the distance in Eqn. (1) becomes the common Euclidean distance.

Definition 1 *The distance metric learning (DML) problem is to learn an optimal distance metric* $\mathbf{A} \in \mathbb{R}^{m\times m}$ *from a collection of data points* $\mathcal{C}$ *on a vector space* $\mathbb{R}^m$ *together with a set of similar pairwise constraints* $\mathcal{S}$ *and a set of dissimilar pairwise constraints* $\mathcal{D}$*, which can be formally formulated into the following optimization framework:*

$$\min_{\mathbf{A}\succeq 0} f(\mathbf{A}, \mathcal{S}, \mathcal{D}, \mathcal{C}) \tag{2}$$

where the metric $\mathbf{A}$ *is a positive semidefinite matrix and* f *is some objective function defined over the given data.*

Given the above definition, the key to solve the DML problem is to formulate a proper objective function f and then find an efficient algorithm to solve the optimization problem. In the following subsections, we will discuss some principles for formulating appropriate optimizations toward DML. We will then emphasize it is important to avoid overfitting when solving a real-world DML problem.

3.2. A Regularization Learning Framework

One common principle for metric learning is to *minimize* the distances between the data points with similar constraints and meanwhile to *maximize* the distances between the data points with dissimilar constraints. We refer it to a **min-max** principle. Some existing DML work can be interpreted within the min-max learning framework. For example, [23] formulated the DML problem as a convex optimization problem:

$$\begin{aligned} \min_{\mathbf{A}\succeq 0} \quad & \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathbf{A}} \\ \text{s.t.} \quad & \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}} \geq 1 \end{aligned} \tag{3}$$

This formulation attempts to find the metric $\mathbf{A}$ by minimizing the sum of squared distances between the similar data points and meanwhile enforcing the sum of distances between the dissimilar data points larger than 1. Although the above method has been shown effective for some clustering tasks, it might not be suitable to solving real-world CIR applications, where the log data could be quite noisy and might be limited at the beginning stage of system development. In practice, the above DML method is likely to overfit the log data in real-world applications.

To enable DML techniques effective for practical applications, the second principle we would like to highlight is the **regularization** principle, which is a key to enhance the generalization and robustness performances of the distance metric in practical applications. Regularization has played a key role in many machine learning methods for preventing overfitting [6]. For example, in SVMs, regularization is critical to ensuring the excellent generalization performance [21].

Similar to the idea of regularization used in kernel machine learning [21], we formulate a general regularization framework for distance metric learning as follows:

$$\begin{aligned} \min_{\mathbf{A}} \quad & g(\mathbf{A}) + \gamma_s \mathcal{V}_s(\mathcal{S}) + \gamma_d \mathcal{V}_d(\mathcal{D}) \\ s.t. \quad & \mathbf{A} \succeq 0 \end{aligned} \tag{4}$$

where $g(\mathbf{A})$ is a regularizer defined on the target metric $\mathbf{A}$, $\mathcal{V}_s(\cdot)$ and $\mathcal{V}_d(\cdot)$ are some loss functions defined on the sets of similar and dissimilar constraints, respectively. γ_s and γ_d are two regularization parameters for balancing the tradeoff between similar and dissimilar constraints as well as the first regularization term. By following the min-max principle, the similar loss function $\mathcal{V}_s(\cdot)$ ($\mathcal{V}_d(\cdot)$) should be defined in the way such that the minimization of the loss function will result in minimizing (maximizing) the distances between the data points with the similar (dissimilar) constraints. In this paper, we adopt the sum of squared distances expression for defining the two loss functions in terms of its effectiveness and efficiency in practice:

$$\mathcal{V}_s(\cdot) = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} \|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathbf{A}} \tag{5}$$

$$\mathcal{V}_d(\cdot) = -\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathbf{A}} \tag{6}$$

In the next subsection, we will discuss how to select an appropriate regularizer and how to incorporate the unlabeled data information via the above regularization learning framework.

3.3. Laplacian Regularized Metric Learning

There are a lot of possible ways to choose a regularizer in the above regularization framework. One simple approach used in [17] is based on the Frobenius norm defined as follows:

$$g(\mathbf{A}) = \|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i,j=1}^{m} a^2_{i,j}} \tag{7}$$

This regularizer simply prevents any elements within the matrix $\mathbf{A}$ from being overlarge. However, the regularizer does not take advantage of any unlabeled data information. In practice, the unlabeled data is beneficial to the DML task. By this consideration, we will show how to formulate a regularizer for exploiting the unlabeled data information in the regularization framework.

Consider the collection of n data points $\mathcal{C}$, we can compute a weight matrix $\mathbf{W}$ between the data points:

$$W_{ij} = \begin{cases} 1 & \mathbf{x}_i \in \mathcal{N}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i) \\ 0 & \text{otherwise.} \end{cases}$$

where $\mathcal{N}(\mathbf{x}_j)$ denotes the nearest neighbor list of the data point $\mathbf{x}_j$. To learn a distance metric, one can assume there is some corresponding linear mapping $\mathbf{U}: \mathbb{R}^m \rightarrow \mathbb{R}^r$, where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \in \mathbb{R}^{m \times r}$, for a possible metric $\mathbf{A}$. As a result, the distance between two input examples can be computed as:

$$\begin{aligned} d(\mathbf{x}_i, \mathbf{x}_j) &= \|\mathbf{U}^\top(\mathbf{x}_i - \mathbf{x}_j)\|^2 \\ &= (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{U}\mathbf{U}^\top (\mathbf{x}_i - \mathbf{x}_j) \\ &= (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{A} (\mathbf{x}_i - \mathbf{x}_j) \qquad (8) \end{aligned}$$

where $\mathbf{A} = \mathbf{U}\mathbf{U}^\top$ is the desirable metric to be learned. By taking unlabeled data information with the weight matrix $\mathbf{W}$, we can formulate the regularizer as follows:

$$\begin{aligned} g(\mathbf{A}) &= \frac{1}{2}\sum_{i,j=1}^{n} \left\|\mathbf{U}^\top \mathbf{x}_i - \mathbf{U}^\top \mathbf{x}_j\right\|^2 W_{ij} \qquad (9) \\ &= \sum_{k=1}^{r} \mathbf{u}_k^\top \mathbf{X}(\mathbf{D} - \mathbf{W})\mathbf{X}^\top \mathbf{u}_k \qquad (10) \\ &= \sum_{k=1}^{r} \mathbf{u}_k^\top \mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{u}_k = \mathbf{tr}(\mathbf{U}^\top \mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{U}) \qquad (11) \\ &= \mathbf{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top) = \mathbf{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{A}) \qquad (12) \end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix whose diagonal elements are equal to the sums of the row entries of $\mathbf{W}$, i.e., $D_{ii} = \sum_j W_{ij}$, and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is known as the Laplacian matrix, and **tr** stands for the *trace* function.

After designing the above Laplacian regularizer, we formulate a new distance metric learning method, called "Laplacian Regularized Metric Learning" (LRML), within the regularization framework as follows:

$$\begin{aligned} \min_{\mathbf{A} \succeq 0} \ \mathbf{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{A}) &+ \gamma_s \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}}^2 \\ &- \gamma_d \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}}^2 \qquad (13) \end{aligned}$$

3.4. LRML Algorithm with Application to CIR

We now show how to apply the proposed LRML technique to collaborative image retrieval and investigate its related optimization in detail. Following the previous work in [13, 17], we assume the log data collected were in the forms of *log sessions*, in which every log session corresponds to a particular user query. In each log session, a user first submits an image example to the CBIR system and then judges the relevance on the top ranked images returned by the CBIR system. The user relevance judgements will then be saved as the log data.

To apply the DML techniques for CIR, for each log session of user relevance feedback, we can convert it into similar and dissimilar pairwise constraints. Specifically, given a specific query q, for any two images $\mathbf{x}_i$ and $\mathbf{x}_j$, if they are marked as relevant in the log session, we will put them into the set of similar pairwise constraints $\mathcal{S}_q$; if one of them is marked as relevant, and the other is marked as irrelevant, we will put them into the set of dissimilar pairwise constraints. As a result, we denote the collection of user relevance feedback log data as $\mathcal{L} = \{(\mathcal{S}_q, \mathcal{D}_q), q = 1, \ldots, Q\}$, where Q is the number of log sessions in the log dataset.

In the CIR context, we can reformulate the two loss functions of the above LRML formulation:

$$\begin{aligned} \min_{\mathbf{A} \succeq 0} \ \mathbf{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{A}) &+ \gamma_s \sum_{q=1}^{Q} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_q} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}}^2 \\ &- \gamma_d \sum_{q=1}^{Q} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}_q} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}}^2 \qquad (14) \end{aligned}$$

To solve the above optimization, we rewrite the two loss functions as follows:

$$\begin{aligned} &\sum_{q=1}^{Q} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_q} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}}^2 \\ &= \sum_{q=1}^{Q} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_q} \mathbf{tr}\Big(\mathbf{A} \cdot (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\Big) \\ &= \mathbf{tr}\left(\mathbf{A} \cdot \sum_{q=1}^{Q} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_q} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\right) \qquad (15) \end{aligned}$$

and

$$\begin{aligned} &\sum_{q=1}^{Q} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}_q} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}}^2 \\ &= \mathbf{tr}\left(\mathbf{A} \cdot \sum_{q=1}^{Q} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}_q} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\right) \qquad (16) \end{aligned}$$

To simplify the above expressions, we introduce two matrices $\mathbf{S}$ and $\mathbf{D}$:

$$\mathbf{S} = \sum_{q=1}^{Q} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_q} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \qquad (17)$$

$$\mathbf{D} = \sum_{q=1}^{Q} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}_q} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \qquad (18)$$

Further, by introducing a slack variable t, we can rewrite the formulation equivalently into the following compact form:

$$\begin{aligned} \min_{\mathbf{A}} \quad & t + \gamma_s \mathbf{tr}(\mathbf{A} \cdot \mathbf{S}) - \gamma_d \mathbf{tr}(\mathbf{A} \cdot \mathbf{D}) \qquad (19) \\ s.t. \quad & \mathbf{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{A}) \leq t \\ & \mathbf{A} \succeq 0 \end{aligned}$$

The above optimization is clearly a standard formulation of Semidefinite Programs (SDP) [2], which can be solved efficiently with global optimum using existing convex optimization packages, such as SeDuMi [19].

4. Experimental Results

In our experiments, we evaluate the effectiveness of LRML for CIR. We design the experiments for performance evaluation in several aspects. First of all, we extensively compare it with a number of state-of-the-art DML techniques. Secondly, we carefully examine if the proposed algorithm is effective to learn reliable metrics by exploiting the unlabeled data for limited log data. Finally, we study if the proposed algorithm is robust to large noisy log data.

4.1. Experimental Testbed

We employ a standard CBIR testbed [13], which consists of 2,000 images in 20 semantic categories from COREL image CDs. Each category consists of exactly 100 images that are randomly selected from relevant examples in COREL CDs. Every category represents a different semantic topic, such as *antelope*, *butterfly*, *cat*, *dog*, and *horse*, etc.

Figure 1. Image examples in the COREL dataset.

4.2. Image Representation

Image representation is a key step for CBIR. Three kinds of features are used to represent the images: color, edge and texture. For color, three types of color moments are extracted: mean, variance and skewness in each color channel (H, S, and V) respectively. A 9-dimensional color moment is used as the color feature. For edge, the edge direction histogram is engaged [13]. An 18-dimensional edge direction histogram is engaged to represent the edge feature. For texture, the Discrete Wavelet Transformation (DWT) is performed on the image with a Daubechies-4 wavelet filter [15]. A 9-dimensional wavelet texture feature is used to describe the texture information. In total, a 36-dimensional feature is used to represent an image.

4.3. Real Log Data of User Relevance Feedback

We obtained the real log data related to the COREL testbed collected by a real CBIR system from the authors in [13]. In their collection, there are two sets of log data. One is a set of normal log data, which contains small noise. The other is a set of noisy log data of relatively large noise. For log data, a *log session* is defined as the basic unit. Each log session corresponds to a regular relevance feedback session, in which 20 images were judged by a user. Thus, each log session contains 20 labeled images that are marked as either "relevant (positive)" or "irrelevant (negative)." Table 1 shows the information of the log data on the two testbeds. More details can be found in [13].

Table 1. The log data collected from users on two datasets

Datasets	Normal Log		Noisy Log	
	#Log Sessions	Noise	# Log Sessions	Noise
20-Cat	100	7.8%	100	16.2%

4.4. Compared Methods and Experimental Setup

We compare the proposed LRML method extensively with two groups of major metric learning techniques: unsupervised approaches and metric learning with side information. We do not compare the DML techniques for supervised classification as they often require explicit class labels, which is unsuitable for CIR. Although it may be unfair to directly compare the unsupervised methods with supervised/semi-supervised metric learning using side information, we still include the unsupervised results. The results could help us examine how effective is the proposed method compared with traditional approaches since there was still limited comprehensive study for applying DML in CIR before. Specifically, the compared schemes include:

- Euclidean: the baseline denoted as "EU" in short.
- Mahalanobis: a standard Mahalanobis metric, denoted as "Mah" in short. Specifically, $\mathbf{A} = \mathbf{P}^{-1}$, where $\mathbf{P}$ is the covariance matrix.
- PCA: classical PCA method [5]. For all unsupervised methods, the number of reduced dimensions r is set to 15 in all experiments.
- MDS: classical Multidimensional Scaling method [3].
- Isomap: an unsupervised method for finding low-dimensional manifold structures with the geometrical information [20].
- LLE: an unsupervised method that computes low-dimensional and neighborhood-preserving embeddings [16].
- Xing: a popular DML method, which solves the DML task with an iterative convex optimization technique [23].
- RCA: Relevance Component Analysis, which learns a linear projection using only equivalent constraints [1].
- DCA: Discriminative Component Analysis, which improves over RCA by engaging dissimilar constraints [11].

Table 2. Average precision of top ranked images on the 20-Category testbed over 2,000 queries with the *normal* log data. For each compared scheme, the first row shows the AP (%) and the second row shows the relative improvement over the baseline (Euclidean) method.

TOP	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	**MAP**
EU	47.88	39.91	35.62	32.73	30.55	28.84	27.53	26.40	25.39	24.44	31.93
Mah	49.40	40.24	35.22	31.52	28.85	26.71	24.94	23.42	22.19	21.09	30.36
	+ 3.2 %	+ 0.8 %	-1.1 %	-3.7 %	-5.6 %	-7.4 %	-9.4 %	-11.3 %	-12.6 %	-13.7 %	-4.9 %
PCA	47.44	39.50	35.33	32.57	30.45	28.76	27.44	26.32	25.35	24.42	31.76
	-0.9 %	-1.0 %	-0.8 %	-0.5 %	-0.3 %	-0.3 %	-0.3 %	-0.3 %	-0.2 %	-0.1 %	-0.5 %
MDS	47.95	39.80	35.69	32.85	30.63	28.90	27.61	26.47	25.47	24.50	31.99
	+ 0.1 %	-0.3 %	+ 0.2 %	+ 0.4 %	+ 0.3 %	+ 0.2 %	+ 0.3 %	+ 0.3 %	+ 0.3 %	+ 0.2 %	+ 0.2 %
LLE	38.58	31.52	28.43	26.26	24.67	23.40	22.34	21.46	20.68	19.87	25.72
	-19.4 %	-21.0 %	-20.2 %	-19.8 %	-19.2 %	-18.9 %	-18.9 %	-18.7 %	-18.6 %	-18.7 %	-19.4 %
Isomap	34.53	27.34	23.74	21.52	20.04	18.92	18.04	17.23	16.56	15.88	21.38
	-27.9 %	-31.5 %	-33.4 %	-34.2 %	-34.4 %	-34.4 %	-34.5 %	-34.7 %	-34.8 %	-35.0 %	-33.0 %
Xing	49.54	42.66	38.88	36.19	34.17	32.51	31.07	29.76	28.61	27.50	35.09
	+ 3.5 %	+ 6.9 %	+ 9.2 %	+ 10.6 %	+ 11.8 %	+ 12.7 %	+ 12.9 %	+ 12.7 %	+ 12.7 %	+ 12.5 %	+ 9.9 %
RCA	51.51	43.16	38.41	35.19	32.70	30.64	29.01	27.56	26.21	24.96	33.94
	+ 7.6 %	+ 8.1 %	+ 7.8 %	+ 7.5 %	+ 7.0 %	+ 6.2 %	+ 5.4 %	+ 4.4 %	+ 3.2 %	+ 2.1 %	+ 6.3 %
DCA	52.63	44.11	39.24	35.95	33.36	31.27	29.58	28.13	26.81	25.51	34.66
	+ 9.9 %	+ 10.5 %	+ 10.2 %	+ 9.8 %	+ 9.2 %	+ 8.4 %	+ 7.4 %	+ 6.6 %	+ 5.6 %	+ 4.4 %	+ 8.6 %
RML	52.09	43.80	39.46	36.37	34.06	32.33	30.74	29.45	28.26	27.20	35.38
	+ 8.8 %	+ 9.7 %	+ 10.8 %	+ 11.1 %	+ 11.5 %	+ 12.1 %	+ 11.7 %	+ 11.6 %	+ 11.3 %	+ 11.3 %	+ 10.8 %
LRML	54.88	46.51	42.03	38.71	36.18	34.05	32.44	30.95	29.66	28.36	37.38
	+ 14.6 %	**+ 16.5 %**	**+ 18.0 %**	**+ 18.3 %**	**+ 18.4 %**	**+ 18.1 %**	**+ 17.8 %**	**+ 17.2 %**	**+ 16.8 %**	**+ 16.0 %**	**+ 17.1 %**

- RML: the regularized metric learning algorithm with the Frobenius norm as the regularizer [17].
- LRML: the proposed Laplacian Regularized Metric Learning algorithm.

In sum, the compared schemes include 2 standard metrics, 3 unsupervised metrics, 4 supervised DML with side information, and the proposed semi-supervised DML method.

For the setup of experiments, we follow a standard procedure for CBIR experiments. Specifically, a query image is picked from the database and then queried with the evaluated distance metric. The retrieval performance is then evaluated based on the top ranked images ranging from top 10 to top 100 images. The average precision (AP) and Mean Average Precision (MAP) are engaged as the performance metrics, which are widely used in CBIR experiments. For the implementation of the proposed LRML algorithm, we use a standard method for computing a normalized Laplacian matrix with 6 nearest neighbors. We fix the two regularization parameters such that γ_d is about one-third of γ_s.

4.5. Experiment I: Normal Log Data

For of all, we evaluate the compared schemes on the normal log data. This is to examine if the proposed algorithm is comparable or better than the previous DML techniques in a normal situation. Table 2 shows the experimental results on the 20-category testbed averaging over 2,000 queries with the normal log data. From the results, we can draw several observations. Firstly, we found that a simple Mahalanobis distance does not always outperform Euclidean distance. In fact, it only improved slightly on top 10 and top 20 ranked images, but failed to obtain improvements on other cases. Secondly, comparing with several unsupervised methods, it is interesting to find that only the MDS method achieved a marginal improvement over the baseline. Two manifold based unsupervised methods performed very poor in this retrieval task. Further, comparing several previous DML methods with the normal log data, the RML method achieved the best overall performance, which obtained 10.8% improvement on MAP over the baseline. The RCA performed the worst among the four compared methods. Finally, comparing with all the metrics, the proposed LRML method achieved the best performance, which significantly improves the baseline with about 17% improvement on MAP. This shows that the proposed method is more effective than the previous methods with normal log data.

4.6. Experiment II: Noisy Log Data

To evaluate the robustness performance, this experiment is to evaluate the performance of the compared schemes with the noisy log data of relatively large noise. Table 3 shows the experimental results on the testbed with the log data of large noise respectively. From the experimental results, we found that the Xing's DML method failed to improve over the baseline method due to the noise problem. The results validated our previous conjecture that the Xing's DML method may be too sensitive to noise. Compared with the Xing's method, the other three DML methods including RCA, DCA and RML are less sensitive to noise, but they still suffered a lot from the noise. Comparing with other approaches, the proposed LRML method is more reliable for achieving a significant improvement. For example, the LRML method achieved 17.1% improvement on MAP with normal log data, and is still able to achieve 14.9% improvement on MAP with the larger noisy log data without dropping too much. These experimental results again validate that the proposed LRML method is effective to learn reli-

Table 3. Average precision (%) of top-ranked images on the 20-Category testbed over 2,000 queries with the *noisy* log data. For each scheme, the first row shows the AP (%) and the second row shows the relative improvement over the baseline (Euclidean) method.

TOP	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	**MAP**
EU	47.88	39.91	35.62	32.73	30.55	28.84	27.53	26.40	25.39	24.44	31.93
Xing	47.90 + 0.0 %	39.87 -0.1 %	35.56 -0.2 %	32.69 -0.1 %	30.52 -0.1 %	28.82 -0.1 %	27.49 -0.1 %	26.37 -0.1 %	25.36 -0.1 %	24.41 -0.1 %	31.90 -0.1 %
RCA	51.14 + 6.8 %	42.59 + 6.7 %	37.75 + 6.0 %	34.45 + 5.3 %	32.00 + 4.7 %	30.00 + 4.0 %	28.31 + 2.8 %	26.97 + 2.2 %	25.69 + 1.2 %	24.45 + 0.0 %	33.34 + 4.4 %
DCA	52.34 + 9.3 %	43.60 + 9.2 %	38.66 + 8.5 %	35.33 + 7.9 %	32.86 + 7.6 %	30.84 + 6.9 %	29.17 + 6.0 %	27.84 + 5.5 %	26.58 + 4.7 %	25.37 + 3.8 %	34.26 + 7.3 %
RML	50.55 + 5.6 %	42.21 + 5.8 %	37.92 + 6.5 %	35.01 + 7.0 %	32.76 + 7.2 %	30.99 + 7.5 %	29.54 + 7.3 %	28.34 + 7.3 %	27.30 + 7.5 %	26.30 + 7.6 %	34.09 + 6.8 %
LRML	53.93 **+ 12.6 %**	45.95 **+ 15.1 %**	41.07 **+ 15.3 %**	37.85 **+ 15.6 %**	35.37 **+ 15.8 %**	33.43 **+ 15.9 %**	31.83 **+ 15.6 %**	30.40 **+ 15.2 %**	29.15 **+ 14.8 %**	27.89 **+ 14.1 %**	36.69 **+ 14.9** %

able distance metrics on real noisy log data by exploiting the unlabeled data information.

5. Conclusion

We proposed a novel semi-supervised distance metric learning scheme for collaborative image retrieval, in which real log data of user relevance feedback were analyzed to discover useful information and infer optimal metrics for image retrieval. To exploit the unlabeled data for the metric learning task, we suggested a new Laplacian Regularized Metric Learning (LRML) algorithm, which leverages the unlabeled data information and ensures metric learning smoothness through a regularization learning framework. We compare the proposed method with a large number of standard options and several new methods proposed recently. The results show that the proposed LRML method is more effective than the state-of-the-art methods for learning reliable metrics from realistic log data that are noisy. In future work, we will conduct more extensive evaluations and investigate more effective techniques to improve the performance.

Acknowledgments

The work was supported in part by Singapore Academic Research Fund Tier 1 Grant (Project No. RG67/07).

References

[1] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *JMLR*, 6:937–965, 2005. 2, 5

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2003. 5

[3] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994. 2, 5

[4] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007. 1

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Elsevier, 1990. 2, 5

[6] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269, 1995. 3

[7] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *NIPS'05*, 2005. 2

[8] X. He, W.-Y. Ma, and H.-J. Zhang. Learning an image manifold for retrieval. In *ACM Multimedia*, pages 17–23, New York, 2004. 2

[9] C.-H. Hoi and M. R. Lyu. Biased support vector machine for relevance feedback in image retrieval. In *Proc. Intl. Joint Conf. on Neural Networks (IJCNN'04)*, Budapest, Hungary, 2004. 1

[10] C.-H. Hoi and M. R. Lyu. Group-based relevance feedback with support vector machine ensembles. In *Proc. 17th International Conf. on Pattern Recognition (ICPR'04)*, volume 3, pages 874–877, Cambridge, UK, 2004. 1

[11] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *Proc. CVPR2006*, New York, US, June 17–22 2006. 2, 5

[12] S. C. H. Hoi, M. R. Lyu, and R. Jin. Integrating user feedback log into relevance feedback by coupled svm for content-based image retrieval. In *Proc. International Conference on Data Engineering Workshops*, 2005. 2

[13] S. C. H. Hoi, M. R. Lyu, and R. Jin. A unified log-based relevance feedback scheme for image retrieval. *IEEE Trans. KDE*, 18(4):509–204, 2006. 1, 2, 4, 5

[14] G. H. J. Goldberger, S. Roweis and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS17*, 2005. 2

[15] B. Manjunath, P. Wu, S. Newsam, and H. Shin. A texture descriptor for browsing and similarity retrieval. *Signal Processing Image Communication*, 2001. 5

[16] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 2, 5

[17] L. Si, R. Jin, S. C. H. Hoi, and M. R. Lyu. Collaborative image retrieval via regularized metric learning. *ACM Multimedia Systems Journal*, 12(1):34–44, 2006. 1, 2, 3, 4, 6

[18] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. PAMI*, 22(12):1349–1380, 2000. 1

[19] J. F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999. 5

[20] J. B. Tenenbaum and V. de Silva andJohn C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 2, 5

[21] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998. 3

[22] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS 18*, pages 1473–1480, 2006. 2

[23] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *NIPS2002*, 2002. 1, 2, 3, 5

[24] L. Yang, R. Jin, R. Sukthankar, and Y. Liu. An efficient algorithm for local distance metric learning. In *AAAI2006*, 2006. 2