Multimodal Graph-Based Reranking for Web Image Search

Meng Wang, *Member, IEEE*, Hao Li, Dacheng Tao, *Senior Member, IEEE*, Ke Lu, and Xindong Wu, *Fellow, IEEE*

***Abstract*— This paper introduces a web image search reranking approach that explores multiple modalities in a graph-based learning scheme. Different from the conventional methods that usually adopt a single modality or integrate multiple modalities into a long feature vector, our approach can effectively integrate the learning of relevance scores, weights of modalities, and the distance metric and its scaling for each modality into a unified scheme. In this way, the effects of different modalities can be adaptively modulated and better reranking performance can be achieved. We conduct experiments on a large dataset that contains more than 1000 queries and 1 million images to evaluate our approach. Experimental results demonstrate that the proposed reranking approach is more robust than using each individual modality, and it also performs better than many existing methods.**

***Index Terms*— Image search, multimodal graph-based learning, reranking.**

I. Introduction

COMMERCIAL image search engines, such as Google[1], Yahoo[2] and Bing[3], usually index web images using textual information, such as images' titles, ALT text and surrounding texts on web pages. However, such text information may not describe the content of images. This fact can severely degrade the search performance of web images.

Several approaches have been investigated to boost the performance of web image search. One approach is image annotation, which aims to associate several keywords to an image to describe its content based on machine learning and computer vision techniques [2]–[4]. However, although great progress has been made in the past few years, automatic annotation of large-scale web images can still hardly achieve satisfactory performance due to the well-known semantic gap. Another approach is web image search reranking. Different from annotation that aims to enhance the text indexing of web images, reranking is applied to directly adjust search results by mining images' visual content [5]–[11].

Most image search reranking methods are developed based on the following two assumptions:

1) The results after reranking should not change too much from the initial ranking list. It means that we assume text information is able to provide a reasonable ranking result.
2) Visually similar images should be close in a ranking list. It is usually called a visual consistency assumption.

A lot of reranking methods are built based on manifold discovery [12], [13]. Manifold-based reranking approach assumes that relevant images lie on a manifold in visual feature space and it is usually accomplished by graph-based learning methods. Therefore, we also call it graph-based reranking. Generally, the approach constructs a graph where the vertices are images and the edges reflect their pairwise similarities. Then, based on the previously mentioned assumptions, a regularization framework is formulated which contains the following two terms: a graph regularizer that keeps the ranking positions of visually similar images close, and a loss term that insures the reranked results do not change too much from the initial ranking list.

Although many different reranking algorithms have been proposed, existing results show that reranking is not guaranteed to improve performance [14]–[17]. In fact, in several cases search performance may even degrade after reranking. One reason can be that the visual consistency assumption does not hold for the employed feature space. The type of the most effective features should vary across queries. For example, for some queries that are related to color distribution, such as sunset, sunrise and beach, color features will be useful. For some queries like building and street, edge and texture features will be more effective. It can be understood that the image

Manuscript received February 3, 2012; revised June 21, 2012; accepted June 22, 2012. Date of publication July 17, 2012; date of current version October 12, 2012. This work was supported in part by the National 863 Program of China under Grant 2012AA011005, the U.S. National Science Foundation (NSF) under Grant CCF-0905337, the National Science Foundation of China under Grant 61103130, Grant 61070120, and Grant 61141014, and the Australian Research Council Discovery under Project DP120103730. A two-page abstract of this paper has been published in ACM SIGIR 2010 [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joan Serra-Sagrista.

M. Wang is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China.

H. Li is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China.

D. Tao is with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering & Information Technology, University of Technology, Sydney, Ultimo, NSW 2007, Australia.

K. Lu is with the Graduate University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: luk@gucas.ac.cn).

X. Wu is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China, and also with the Department of Computer Science, University of Vermont, Burlington, VT 05405 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2207397

[1]Available at http://images.google.com/.

[2]Available at http://images.search.yahoo.com/.

[3]Available at http://www.bing.com/images/.

manifolds for these queries exhibit in different feature spaces [12]. Therefore, employing multimodal features can be a solution. Note that multiple modalities are frequently used to denote different types of media data, such as image and text. But here a modality is viewed as a description of image data, such as color, edge and texture. In our work, it can be used with "feature set" interchangeably. Thus, employing multimodal features means exploring multiple visual feature sets instead of combining visual and textual information. Using multimodal features can guarantee that the useful features for different queries are contained, but there are still several problems that need to be addressed, such as how to adaptively integrate different modalities and discover the most useful modalities.

Early fusion and late fusion are the two most popular approaches for using multimodal features [18]. Early fusion means concatenating multimodal features into a long feature vector, and late fusion integrates the results obtained by learning with each modality. But the early fusion approach usually suffers from the "curse-of-dimensionality" problem. For late fusion, the fused results may not be good since each modality might be poor. In addition, it will be difficult to assign appropriate weights to different modalities. Therefore, in this work we propose a multimodal graph-based learning approach that can adaptively integrate multiple modalities. We simultaneously integrate the learning of relevance scores, weights of modalities, and the distance metric and its scaling of each modality (the scaling is used to estimate the similarity of sample pairs in a modality) into a unified graph-based learning scheme. Via adaptively modulating the weights of different modalities, the proposed scheme is able to optimally integrate these modalities for reranking. Figure 1 illustrates the web image search reranking scheme based on the approach. We conduct experiments on a large dataset that contains more than 1,000 queries and 1 million images to evaluate our approach. Experimental results demonstrate that the proposed reranking approach is much more robust than using each individual modality. It also shows averagely better performance than many other methods.

The contribution of this work is summarized as follows:

1) We propose a multimodal graph-based learning approach for web image search reranking. It is able to integrate multiple modalities into a graph-based learning framework.
2) The proposed approach simultaneously learns the relevance scores, weights of modalities, and the distance metric and its scaling for each modality. Although multiple modalities are involved, there are only two parameters in our algorithm and this makes the approach robust and flexible.
3) We conduct an empirical study on more than 1,000 queries and 1 million images. This compares favorably than many existing works on reranking that conduct experiments on small datasets.

The rest of this paper is organized as follows. In Section II, we introduce related work, including visual search reranking, multimodal fusion and graph-based learning. In Section III, we introduce the formulation and the solution of our algorithm.

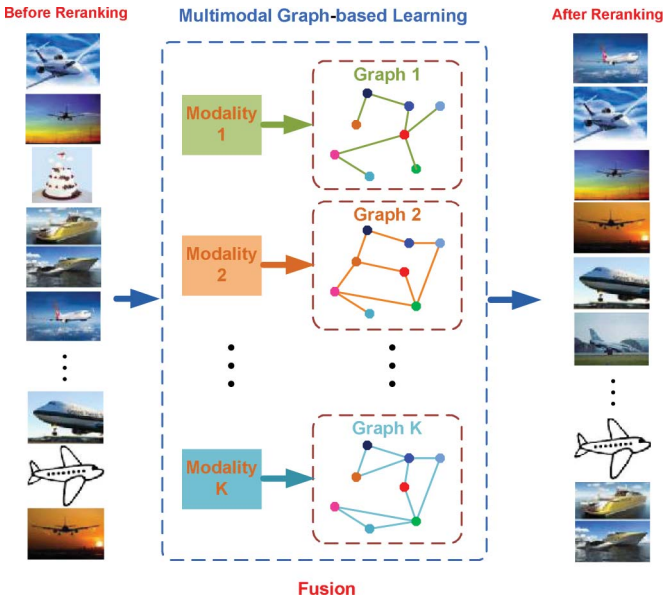


Fig. 1. Schematic illustration of the web image search reranking approach based on multimodal graph-based learning.

In Section IV, we introduce experiments, including experimental settings, experimental results and discussion. Finally, we conclude the paper in Section V.

II. Related Work

A. Visual Search Reranking

Visual search reranking has been widely investigated for improving the search performance of web images, photos and other multimedia documents. The existing visual search reranking efforts can mainly be classified into two categories according to whether there are query examples available. For the first category, which can be named example-based reranking, there are several examples along with a text query. Yan et al. [19] regard the query examples as pseudo relevant samples and collect several bottom results in a ranking list as pseudo irrelevant ones. An SVM model is learned based on these samples to rerank search results. Natsev et al. [20] improve the robustness of this approach by a bagging strategy. They collect multiple pseudo irrelevant sample sets and then generate different ranking lists accordingly. These ranking lists are aggregated to generate final results. Liu et al. [21] identify an optimal set of document pairs via an information theory principle and a ranking list is directly recovered from this pair set. These methods can effectively improve search performance if good visual examples are provided. But they cannot be used in the cases when there is no visual example available.

The other approach does not rely on query examples. It aims to improve text-based search by mining the visual information of images or videos. Kennedy and Chang [22] regard top and bottom results in a ranking list as pseudo relevant and irrelevant samples respectively to discover the related concepts. The detection results of the related concepts are then used as high-level features in SVM to build classifiers for reranking. Hsu et al. [14] formulate the reranking process as a random walk over a context graph, where video stories are nodes

and the edges between them are weighted by multimodal similarities. Jing et al. [23] employ a random walk process to rerank Google image search results by mining the visual similarity of search results. Tian et al. [17] propose a graph-based approach, which encodes the assumptions that the reranked results do not change much from the initial ranking list and the ranking positions of visually similar images are close. Yang et al. [24] extract multiple features from each image and collect a training set that contains several queries and labeled search results. Reranking is then regarded as a supervised learning task. Yao et al. [25] propose a method that can simultaneously explore the visual content and textual information of web images. Several methods have shown effectiveness in standard challenges, such as ImageCLEF photo search and Wikipedia retrieval tasks [26]–[28]. However, these methods all rely on the adopted feature space. They will not work well if the features cannot effectively describe the query semantics. In this work, we investigate image search reranking with multiple modalities that describe images' visual content from different aspects. By adaptively learning the weighting parameters, we will show that our approach can effectively integrate multiple modalities to boost ranking performance.

B. Multimodal Visual Feature Fusion

Existing studies reveal that the distances between samples become increasingly similar when the dimension of adopted feature space is high. This may introduce performance degradation if we directly apply high-dimensional features to distance (or similarity)-based learning algorithms, such as the graph-based method adopted in this work. To deal with this issue, a natural method is to replace the high-dimensional learning task by multiple low-dimensional learning tasks, i.e., separately applying different modalities to learning algorithms and then fusing the results [29]. A modality can be viewed as a description to image or video data, such as color, edge, texture, audio, and text. This method is usually called "multimodal fusion" or "multimodality learning". Sometimes it is also named "late fusion", whereas the approach of using concatenated high-dimensional global feature vector is named "early fusion" [18]. With a labeled fusion set, the task can actually be formulated as a learning issue. For example, Iyengar et al. [30] and Snoek et al. [18] accomplish the fusion with Support Vector Machine (SVM) models. Yan et al. have studied the theoretical upper bound of linear fusion [31]. Snoek et al. provide an empirical study to compare early fusion and late fusion [18]. However, the early and late fusion approaches have their own disadvantages, such as the "curse of dimensionality" in early fusion and the difficulty in determining appropriate weights for late fusion. Wang et al. [32] propose an approach to integrate the graph representations generated from multiple modalities for video annotation. Geng et al. [33] integrate the graph representations in a kernelized learning approach. But these methods cannot be applied to image search reranking and they also fail to discover an appropriate distance metric for each modality.

Our work integrates multiple modalities into a graph-based learning algorithm for reranking. In addition to the learning of images' relevance scores of and the weights of different modalities, our approach further learns the distance metric for each modality and its scaling, and this makes our method more effective and flexible.

C. Graph-Based Learning

Graph-based learning methods have attracted great research interests in the past years [34], [35]. In these methods, a graph is constructed based on the given data, where vertices are samples and edges reflect their similarities. They are usually formulated in a regularization scheme with two terms. One term is used to enforce the function to be smooth on the graph and the other term is used to keep the function consistent with prior information, such as the labeling information of several samples. The algorithms can also be accomplished by a random walk process. In [36], He et al. adopt a graph-based method named manifold-ranking in image retrieval, and Yuan et al. [37] then apply the same algorithm to video annotation. Wang et al. develop a multi-graph learning method for video annotation [32]. Several different graph-based learning approaches have been investigated for reranking [14], [23], [17] (in the next section we will introduce the details). However, there is no investigation of integrating multiple modalities in graph-based learning for reranking. We will show that our approach can achieve better performance by adaptively learning the integration of multiple modalities and the distance metric of each modality.

III. Web Image Search Reranking With Multimodal Graph-Based Learning

In this section, we describe our reranking approach. We first introduce the existing graph-based reranking methods with a general regularization scheme. We then present our approach, including initial relevance score estimation and multimodal graph regularization. We also provide a probabilistic explanation on our formulation, and after that we detail our solution of the optimization problem. For clarity, we illustrate important notations and definitions throughout this paper in Table I.

A. Graph-Based Reranking

We first follow [17] to define several terms in reranking.

Definition 1: A ranking score list, $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$, is a vector of ranking scores, which corresponds to a sample set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$.

Reranking aims to obtain a new ranking score list by performing learning based on images' visual content.

Definition 2: A reranking function h is defined as

$$\mathbf{y} = h(\mathbf{X}, \bar{\mathbf{y}}) \tag{1}$$

where $\bar{\mathbf{y}} = [\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n]^T$ is the initial ranking score list.

Generally, graph-based reranking can be formulated as a regularization framework as follows

$$\text{minimize} \quad Q(\mathbf{y}, \bar{\mathbf{y}}, \mathcal{X}) = R(\mathbf{y}, \mathcal{X}) + \lambda L(\mathbf{y}, \bar{\mathbf{y}}) \tag{2}$$

Here $R(.)$ is a regularization term that makes the ranking scores of visually similar images close, the term $L(.)$ is a loss

TABLE I
NOTATIONS AND DEFINITIONS

Notation	Definition
$\mathbf{x}_i$	The ith image in a reranking task.
$\mathbf{x}_{k,i}$	The kth modality of the ith image. That means $\mathbf{x}_i = \left[\mathbf{x}_{1,i}^T, \mathbf{x}_{2,i}^T, \ldots, \mathbf{x}_{K,i}^T\right]^T$.
$\mathcal{X}$	$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. It is the image set in the reranking task.
$\bar{y}_i$	The initial ranking score of $\mathbf{x}_i$.
$\bar{\mathbf{y}}$	$\bar{\mathbf{y}} = [\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n]^T$. It is the vector of the initial ranking scores.
y_i	The relevance score of $\mathbf{x}_i$, which needs to be estimated in reranking.
$\mathbf{y}$	$\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$. It is the vector of the relevance scores.
$\mathbf{W}_k$	The similarity matrix of images for the kth modality. Its (i, j)th element indicates the similarity of $\mathbf{x}_i$ and $\mathbf{x}_j$ in the kth modality.
$\widetilde{\mathbf{L}}_k$	The normalized graph Laplacian derived from the kth modality.
$\mathbf{A}_k$	The transformation matrix for the kth modality.
d_k	The dimensionality of the kth modality.
α	The weight vector which is used to combine the K normalized graph Laplacians.
λ, ξ	Positive parameters used to modulate the effects of regularizers [see (11)].
n	The number of images.
K	The number of modalities.
N	The neighborhood size for sparsifying similarity matrices.
T	The iteration time in the alternating optimization process for solving (11).
T_1	The iteration time in the gradient descent process for solving $\mathbf{A}_k$ (see Algorithm 1).
T_2	The iteration time in the coordinate descent process for solving α (see Section III-E3).

term that estimates the difference between $\mathbf{y}$ and $\bar{\mathbf{y}}$, and $\mathbf{W}$ is a similarity matrix in which W_{ij} indicates the visual similarity of $\mathbf{x}_i$ and $\mathbf{x}_j$. The term $R(r, \mathcal{X}, \mathbf{W})$ usually employs one of the following two forms:

1) Graph Laplacian regularizer, i.e.,

$$R(\mathbf{y}, \mathcal{X}) = \sum_{i,j} W_{ij}(y_i - y_j)^2 = \mathbf{y}^T \mathbf{L} \mathbf{y} \tag{3}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is called graph Laplacian. Here $\mathbf{D}$ is a diagonal matrix and its (i, i)-th element is the sum of the i-th row of $\mathbf{W}$.

2) Normalized graph Laplacian regularizer, i.e.,

$$R(\mathbf{y}, \mathcal{X}) = \sum_{i,j} W_{ij} \left(\frac{y_i}{d_{ii}} - \frac{y_j}{d_{jj}}\right)^2 = \mathbf{y}^T \widetilde{\mathbf{L}} \mathbf{y} \tag{4}$$

where d_{ii} is the sum of the i-th row of $\mathbf{W}$, and $\widetilde{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ is named normalized graph Laplacian.

For the loss term, usually it estimates the difference between two ranking lists. It can be defined based on either the relevance scores or ranking scores. There are several different choices, such as the squared difference and the hinge distance of relevance scores. More details and discussion on the distance of ranking lists for reranking can be found in [17].

B. Proposed Multimodal Graph-Based Reranking Algorithm

We develop our approach based on normalized graph Laplacian and squared loss. We choose normalized graph Laplacian because existing studies have demonstrated its effectiveness over graph Laplacian [17], and squared loss term is used because it can make the optimization framework easy to solve. First, we present the formulation with one modality. Next, we extend it to multiple modalities.

Typically, the similarity between the i-th and the j-th samples is estimated based on

$$W_{ij} = \exp\left(-\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \tag{5}$$

where σ is the radius parameter of a Gaussian function that converts distance to similarity. However, Euclidean distance may not be appropriate as the most suitable distance metric usually relies on feature distribution [38]–[41]. Therefore, we replace the Euclidean distance metric with a Mahalanobis distance metric in Eq. (5) which can be learned by an optimization framework. The equation thus turns to

$$W_{ij} = \exp\left(-(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)\right) \tag{6}$$

where $\mathbf{M}$ is a symmetric positive semi-definite real matrix. We decompose $\mathbf{M}$ as $\mathbf{M} = \mathbf{A}^T\mathbf{A}$, where $\mathbf{A}$ is a d-by-d matrix. We substitute it into Eq. (6). The equation then becomes

$$W_{ij} = \exp\left(-\|\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)\|^2\right) \tag{7}$$

Actually it is equivalent to transforming each sample $\mathbf{x}$ to $\mathbf{Ax}$. That is, we assume that the manifold of images can be better discovered in a transformed space.

Now we consider there are K modalities. Here we linearly combine the normalized graph Laplacian regularizers

generated from different modalities

$$R(\mathbf{y}, \mathcal{X}, \alpha, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K) = \mathbf{y}^T \widetilde{\mathbf{L}}_k \mathbf{y} = \sum_{k=1}^{K} \sum_{ij} \alpha_k W_{k,ij} \left(\frac{y_i}{d_{k,ii}} - \frac{y_j}{d_{k,jj}} \right)^2 \quad (8)$$

where

$$W_{k,ij} = \exp\left(-\|\mathbf{A}_k(\mathbf{x}_{k,i} - \mathbf{x}_{k,j})\|^2 \right) \quad (9)$$

and α_k is the weight for the k-th modality. The weights satisfy $0 \le \alpha_k \le 1$ and $\sum_{k=1}^{K} \alpha_k = 1$.

As previously mentioned, we integrate the learning of the weights into our regularization framework in order to adaptively modulate the impacts of different modalities. Therefore, the regularizer term turns to

$$R(\mathbf{y}, \mathcal{X}, \alpha, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K) = \sum_{k=1}^{K} \sum_{i,j} \alpha_k W_{k,ij} \left(\frac{y_i}{d_{k,ii}} - \frac{y_j}{d_{k,jj}} \right)^2 + \xi \|\alpha\|^2 \quad (10)$$

Accordingly, our algorithm can be formulated as the following optimization problem

$$\min_{\mathbf{y}, \alpha, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K} Q(\mathbf{y}, \mathcal{X}, \alpha, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K) = \sum_{k=1}^{K} \sum_{i,j} \alpha_k W_{k,ij} \left(\frac{y_i}{d_{k,ii}} - \frac{y_j}{d_{k,jj}} \right)^2 + \lambda \|\mathbf{y} - \bar{\mathbf{y}}\|^2 + \xi \|\alpha\|^2$$

$$\text{s.t.} \quad 0 \le \alpha_k \le 1, \sum_{k=1}^{K} \alpha_k = 1 \quad (11)$$

We can see that we need to solve the following variables: (1) $\mathbf{y}$, i.e., the ranking scores to be estimated; (2) α_k, i.e., the weights for combining the K modalities; and (3) $\mathbf{A}_k, (1 \le k \le K)$, i.e., the transform matrices for the K modalities. Note that an appropriate scale of $\mathbf{A}_k$ for estimating $\mathbf{W}_k$ will also be automatically determined, as there is no radius parameter in Eq. (7). The radius parameter is usually very sensitive for graph-based learning and it needs to be carefully tuned [42], [43]. The elimination of the radius parameter by automatically determining the scale of $\mathbf{A}_k$ is also a benefit of our approach.

We first introduce the estimation of initial relevance scores $\bar{\mathbf{y}}$ and the probabilistic explanation of our approach, and the solution of the optimization problem will be explained later.

C. Initial Relevance Estimation

Since in reranking we only have original ranking lists instead of quantized scores, a necessary step is to turn the ranking positions into scores. Traditional methods usually associate $\bar{y}_i$ with the position τ_i using heuristic strategies, such as $\bar{y}_i = 1 - \frac{\tau_i}{n}$ or $\bar{y}_i = n - \tau_i$. In this work, we investigate the relationship between $\bar{y}_i$ and the position τ_i with a large number of queries. Actually, we can define

$$\bar{y}_i = E_{q \in \mathcal{Q}}[\hat{y}(q, \tau_i)] \quad (12)$$

where $\mathcal{Q}$ means the set of all possible queries, $E_{q \in \mathcal{Q}}$ means the expectation over the query set $\mathcal{Q}$, and $\hat{y}(q, \tau_i)$ indicates the relevance ground truth of the i-th search result for query q. Therefore, the most intuitive approach is to estimate $\bar{y}_i$ by averaging $\hat{y}(q, \tau_i)$ over a large query set. Figure 2 illustrates the results obtained by using more than 1,000 queries. Here the relevance score of each search result is manually labeled to be 0, 1, or 2. Details about the queries and the dataset will be introduced in Section IV. However, as shown in Fig. 2, the averaged relevance score curve with respect to the ranking position is not smooth enough even after using more than 1,000 queries. A prior knowledge can be that the expected relevance score should be decreasing with respect to ranking position. Therefore, we further smooth the curve with a parametric approach. We assume $\bar{y}_i = a + be^{-i/c}$ and then fit this function with the non-smooth curve. In this way, we estimate the parameters a, b, and c with mean squared loss criterion. The values of a, b, and c are estimated to be 1.208, 0.4266, and 141.22, respectively. Figure 2 shows the fitted curve, and we can see that it reasonably preserves the original information.

D. Probabilistic Explanation

Now we provide a probabilistic explanation for our approach. From a probabilistic perspective, we can derive the optimal $\mathbf{y}, \alpha, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K$ with the maximum posterior probability given the samples $\mathcal{X}$ and initial relevance scores $\bar{\mathbf{y}}$

$$\{\mathbf{y}, \alpha, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K\}^* = \arg\max p(\mathbf{y}, \alpha, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K | \mathcal{X}, \bar{\mathbf{y}}) \quad (13)$$

Following Bayes rule, the above equation can turn to

$$\arg\max p(\mathbf{y}, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K | \mathcal{X}, \alpha) p(\bar{\mathbf{y}} | \mathcal{X}, \mathbf{y}, \alpha) p(\alpha) \quad (14)$$

We let

$$p(\mathbf{y}, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K | \mathcal{X}, \alpha) = \frac{1}{Z_1} \exp\left(-\sum_{k=1}^{K} \sum_{i,j} \alpha_k W_{k,ij} \left(\frac{y_i}{d_{k,ii}} - \frac{y_j}{d_{k,jj}} \right)^2 \right) \quad (15)$$

$$p(\bar{\mathbf{y}} | \mathcal{X}, \mathbf{y}, \alpha) = p(\bar{\mathbf{y}} | \mathcal{X}, \mathbf{y}) = \frac{1}{Z_2} \exp\left(-\frac{\|\mathbf{y} - \bar{\mathbf{y}}\|^2}{1/\lambda} \right) \quad (16)$$

and

$$p(\alpha) = \frac{1}{Z_3} \exp\left(-\frac{\|\alpha - \frac{1}{K}\mathbf{1}\|^2}{1/\xi} \right) \quad (17)$$

where Z_1, Z_2 and Z_3 are normalizing constants which keep the integral of the probability function to be 1, and $\mathbf{1}$ is a vector that has all the entries to be 1. The first two terms have been explained in [17]. In comparison with the probabilistic scheme in [17], we integrate K normalized graph Laplacians in the terms and add the third term. The third term actually adds a Gaussian distribution prior to α, and its mean vector is $\frac{\mathbf{I}}{K}$, i.e., an average prior. By adding a constraint $\sum_{k=1}^{K} \alpha_k = 1$, we then see that Eq. (14) and Eq. (11) are equivalent.

E. Solution

We adopt an alternating optimization to solve Eq. (11). More specifically, we alternatively update $\mathbf{y}$, α, and $\mathbf{A}_k (k = 1, 2, \ldots, K)$ to optimize the objective.

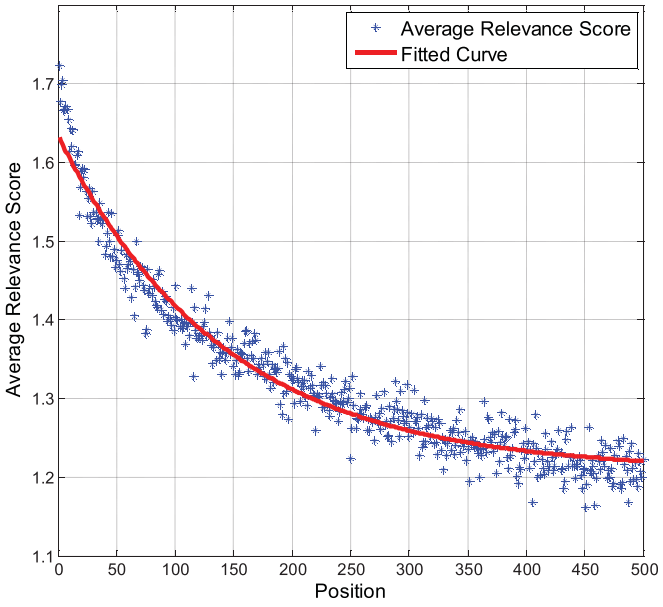


Fig. 2. Average relevance scores at different ranking positions.

1) Optimization of y: We fix α and $\mathbf{A}_k (k = 1, 2, \ldots, K)$, and then we can easily derive that

$$\mathbf{y} = \left(\mathbf{I} + \frac{1}{\lambda} \sum_{k=1}^{K} \alpha_k \widetilde{\mathbf{L}}_k \right)^{-1} \bar{\mathbf{y}} \tag{18}$$

We can see that, in comparison with general normalized graph Laplacian based learning, the only difference is that the K normalized graph Laplacian matrices have been linearly combined with weights α_k.

2) Optimization of A_k: Now we consider the optimization of $\mathbf{A}_k$. Considering $\mathbf{y}$, α, and $\mathbf{A}_1, \mathbf{A}_2, \ldots \mathbf{A}_{k-1}, \mathbf{A}_{k+1}, \ldots \mathbf{A}_k$ are fixed, then we derive the derivative of Q with respect to $\mathbf{A}_k$ as

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{A}_k} Q(\mathbf{y}, \mathcal{X}, \alpha, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K) \\ &\quad = \alpha_k \frac{\partial}{\partial \mathbf{A}_k} \sum_{i,j} W_{k,ij} \left(\frac{y_i}{d_{k,ii}} - \frac{y_j}{d_{k,jj}} \right)^2 \\ &\quad = \alpha_k \sum_{ij} h_{ij}^2 \frac{\partial W_{k,ij}}{\partial \mathbf{A}_k} - W_{k,ij}^T h_{ij} \\ &\qquad \times \left(\frac{y_i}{\sqrt{d_{k,ii}^3}} \frac{\partial d_{k,ii}}{\partial \mathbf{A}_k} - \frac{y_j}{\sqrt{d_{k,jj}^3}} \frac{\partial d_{k,jj}}{\partial \mathbf{A}_k} \right) \end{aligned} \tag{19}$$

where

$$h_{ij} = \frac{y_i}{d_{k,ii}} - \frac{y_j}{d_{k,jj}} \tag{20}$$

$$\frac{\partial W_{k,ij}}{\partial \mathbf{A}_k} = -2 W_{k,ij} \mathbf{A}_k (\mathbf{x}_{k,i} - \mathbf{x}_{k,j})^T (\mathbf{x}_{k,i} - \mathbf{x}_{k,j}) \tag{21}$$

$$\frac{\partial d_{k,ii}}{\partial \mathbf{A}_k} = \sum_{j=1}^{n} \frac{\partial W_{k,ij}}{\partial \mathbf{A}_k} \tag{22}$$

Based on the derivative, we adopt a gradient descent process to solve the optimization of $\mathbf{A}_k$.

In the gradient descent process, we dynamically adapt the step-size in order to accelerate the process while guaranteeing its convergence. Denote by $\mathbf{A}_k^{(t)}$ the values of $\mathbf{A}_k$ in the t-th turn of the iterative process. If $Q(\mathbf{y}, \mathcal{X}, \alpha, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k^{(t+1)}, \ldots, \mathbf{A}_K) < Q(\mathbf{y}, \mathcal{X}, \alpha, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k^{(t)}, \ldots, \mathbf{A}_K)$, i.e., the cost function obtained after gradient descent is reduced, then we double the step-size; otherwise, we decrease the step-size and do not update $\mathbf{A}_k$, i.e., $\mathbf{A}_k^{(t+1)} = \mathbf{A}_k^{(t)}$. The process is illustrated in Algorithm 1. In this process, other variables besides $\mathbf{A}_k$ are fixed, and thus we use $Q(\mathbf{A}_k^{(t)})$ to denote the value of the objective function for simplicity.

3) Optimization of α: Considering $\mathbf{y}$ and $\mathbf{A}_k (k = 1, 2, \ldots, K)$ are fixed, then Eq. (11) becomes

$$\begin{aligned} &\min_{\alpha} \sum_{k=1}^{K} \alpha_k g_k + \xi \|\alpha\|^2 \\ &\text{s. t.} \quad 0 \le \alpha_k \le 1, \sum_{k=1}^{K} \alpha_k = 1 \end{aligned} \tag{23}$$

where $g_k = \sum_{i,j} W_{k,ij} \left(\frac{y_i}{d_{k,ii}} - \frac{y_j}{d_{k,jj}} \right)^2 = \mathbf{y}^T \widetilde{\mathbf{L}}_k \mathbf{y}$.

We adopt a coordinate descent method to solve Eq. (23). In each iteration, we select two elements to update and fix others. Suppose the i-th and the j-th elements are selected. Since $\sum_{k=1}^{K} \alpha_k = 1$, $\alpha_i + \alpha_j$ will not change in the process. Therefore, the updating will follow the rule of

$$\begin{cases} \alpha_i^* = 0, \alpha_j^* = \alpha_i + \alpha_j, & \text{if} \quad 2\xi(g_i + g_j) + (\alpha_j - \alpha_i) \le 0 \\ \alpha_i^* = \alpha_i + \alpha_j, \alpha_j^* = 0, & \text{if} \quad 2\xi(g_i + g_j) + (\alpha_i - \alpha_j) \le 0 \\ \alpha_i^* = \frac{2\xi(g_i + g_j) + (\alpha_j - \alpha_i)}{4\xi}, \alpha_j^* = \alpha_i + \alpha_j - \alpha_i^*, & \text{otherwise} \end{cases} \tag{24}$$

We iterate the process for all pairs of elements in α. Since the objective of Eq. (23) will not increase for each step, the process is guaranteed to converge. Note that $g_k = \mathbf{y}^T \widetilde{\mathbf{L}}_k \mathbf{y}$ actually measures the consistency of relevance scores and visual similarity, and thus a smaller value of g_k indicates the smoothness of relevance scores in the k-th modality. Therefore, Eq. (24) indicates that the modality in which the manifold is more consistent with the relevance scores will be strengthen. In addition, due to the constraints $0 \le \alpha_k \le 1$ and $\sum_{k=1}^{K} \alpha_k = 1$, several weights will be 0, i.e., our approach not only adaptively integrates multiple modalities but also has certain ability of feature selection.

The whole alternating optimization process is illustrated in Algorithm 2. Since the objective is lower bounded by 0 and it will keep decreasing in each step, its convergence is guaranteed.

F. Computational Cost

From the above solution process, we can see that its computational cost mainly contains three parts, which are for updating $\mathbf{y}$, $\mathbf{A}_k (k = 1, 2, \ldots, K)$ and α, respectively. From Eq. (18) we can see that the cost for updating $\mathbf{y}$ scales as $O(n^3)$. For updating $\mathbf{A}_k$, from the process in Section III-E 2) we can see that the cost scales as $O(T_1 n^2 d_k^2)$. For updating α, the cost scales as $O(T_2 K^2)$. Therefore, the cost of the whole solution process scales as $O(T(n^3 + T_1 n^2 \sum_{k=1}^{K} d_k^2 + T_2 K^2))$, where n

Algorithm 1 Gradient Descent Process for Solving $\mathbf{A}_k$

Step 1:
1.1: Set t to 0.
1.2: Set step-size parameter η_t to 1.
1.3: Set $\mathbf{A}_k^{(t)}$ to a diagonal matrix $\mathbf{I}/\sigma$. Here σ is determined to be the one in the set $\{\sigma_0/8, \sigma_0/4, \sigma_0/2, \sigma_0, 2\sigma_0, 4\sigma_0, 8\sigma 0\}$ that yields the minimum cost Q, where σ_0 is the median value of the pairwise Euclidean distances.

Step 2:
Let $\mathbf{A}_k^{(t+1)} = \mathbf{A}_k^{(t)} - \eta_t \frac{\partial Q}{\partial \mathbf{A}_k}|_{\mathbf{A}_k=\mathbf{A}_k^{(t)}}$.

Step 3:
If $Q(\mathbf{A}_k^{(t+1)}) < Q(\mathbf{A}_k^{(t)})$, $\eta_{t+1} = 2\eta_t$;
otherwise, $\mathbf{A}_k^{(t+1)} = \mathbf{A}_k^{(t)}$, $\eta_{t+1} = \eta_t/2$.

Step 4:
Let $t = t + 1$. If $t > T_1$, quit iteration and output $\mathbf{A}_k$, otherwise go to step 2.

is the number of samples, d_k is the dimensionality of the k-th modality, K is the number of modalities, and T, T_1 and T_2 are the iteration times of alternating optimization, the gradient descent process in Algorithm 1 and the coordinate descent method for updating α, respectively. To further reduce the computational cost, here we adopt a strategy. We sparsify $\mathbf{W}_k$ by only keeping the N largest components in each row. This is a widely-applied strategy in graph-based learning for reducing computational cost while maintaining performance [42], [44]. For Eq. (18), we adopt an iterative method to solve it instead of using matrix inverse, which is analogous to the method in [34]. In this way, the computational costs for updating $\mathbf{y}$ and $\mathbf{A}_k$ become $O(nN)$ and $O(T_1 nNd_k^2)$, respectively. Therefore, the overall computational cost is $O(T(T_1 nN\sum_{k=1}^{K} d_k^2 + T_2 K^2))$.

Here we also analyze the computational costs of several other reranking methods for comparison. For the random walk method proposed in [23] and the Bayesian reranking method proposed in [17], it can be analyzed that the computational costs are $O(n^2 d + n^3)$. In practice, the time costs are close to our method if the values of n and d are not large. Then, we consider the early fusion method that directly concatenates all the features into a long feature vector. Its computational cost will be $O(TT_1 nNd^2)$. Since $\sum_{k=1}^{K} d_k^2$ is much smaller than d^2, our method is more computationally efficient.

IV. Experiments

In this section, we first introduce our experimental settings, and then we present the experimental results that validate the effectiveness of our approach. The experiments actually contain two parts. In the first part, we will compare our approach with those methods that only use a single modality. In the second part, we compare our algorithm with several existing methods that adopt all features.

A. Experimental Settings

To empirically evaluate the proposed approach, we conduct experiments on the MSRA-MM Version 2.0 dataset [45], which contains 1097 queries. The queries are obtained from a query log of a commercial search engine and they are mainly

Algorithm 2 Alternating Optimization Process of the Proposed Reranking Algorithm

Step 1: Initialization.
1.1: Set t to 0.
1.2: Set $\mathbf{A}_1^{(t)}, \mathbf{A}_2^{(t)}, \ldots, \mathbf{A}_K^{(t)}$ to diagonal matrices $\frac{\mathbf{I}}{\sigma_1}, \frac{\mathbf{I}}{\sigma_2}, \ldots, \frac{\mathbf{I}}{\sigma_K}$, respectively, where σ_k is the median value of the pairwise Euclidean distances of the samples in the k-th modality.
1.3: Construct the similarity matrices $\mathbf{W}_1^{(t)}, \mathbf{W}_2^{(t)}, \ldots, \mathbf{W}_K^{(t)}$.
1.4: Compute $\mathbf{D}_1^{(t)}, \mathbf{D}_2^{(t)}, \ldots, \mathbf{D}_K^{(t)}$ and $\widetilde{\mathbf{L}}_1^{(t)}, \widetilde{\mathbf{L}}_2^{(t)}, \ldots, \widetilde{\mathbf{L}}_K^{(t)}$ accordingly.

Step 2: Relevance score update. Compute the optimal $\mathbf{y}$ according to Eq. (18), i.e.,

$$\mathbf{y}^{(t)} = \left(\mathbf{I} + \frac{1}{\lambda}\sum_{k=1}^{K} \alpha_k \widetilde{\mathbf{L}}_k^{(t)}\right)^{-1} \bar{\mathbf{y}}$$

Step 3: Distance metric update. Update $\mathbf{A}_1^{(t+1)}, \mathbf{A}_2^{(t+1)}, \ldots, \mathbf{A}_k^{(t+1)}$ sequentially by Algorithm 1.

Step 4: Update the weights according to Eq. (24).

Step 5: After obtaining $\mathbf{A}_k^{(t+1)}$, update the similarity matrices $\mathbf{W}_k^{(t+1)}$ with the entries computed as Eq. (9). Then, compute $\mathbf{D}^{(t+1)}$ and $\widetilde{\mathbf{L}}_k^{(t+1)}$ accordingly.

Step 6: Let $t = t + 1$. If $t > T$, quit iteration and output the relevance scores; otherwise, go to step 2.

hot queries that appear most frequently. In [45], the queries have been manually classified into 9 categories. Table II illustrates the number of queries for each category and several examples. We choose this dataset to evaluate our approach for the following reasons: 1) it is a real-world web image dataset; 2) it contains the original ranking information of a popular search engine, and thus we can easily evaluate whether our approach can improve the performance of the search engine; 3) it is publicly available; and 4) it contains more than 1,000 queries that cover widely. For each query, up to 1000 image search results have been collected in the dataset, and there are 1,011,738 images in total. Each image is labeled with a 3-level relevance, i.e., very relevant, relevant and irrelevant. We use scores 2, 1 and 0 to indicate the three relevance levels, respectively. The ambiguity of queries has also been taken into consideration in the labeling process. For example, if a query has multiple semantics, then an image will be labeled as relevant if it is consistent with one of the semantics. More details about the labeling process can be found in [45]. Figure 3 illustrates several example images of "Barak Obama", "Butterfly" and "ipod" with different relevance levels.

There are 7 global features extracted, including

1) 225-dimensional Block-wise color moments. Each image is split into 5-by-5 blocks, and 9-dimensional color moment features are extracted from each block.
2) 64-dimensional HSV color histogram. A 64-dimensional histogram feature vector is extracted in HSV color space for each image.
3) 144-dimensional Color autocorrelogram. HSV color moments are quantized into 36 bins with 4 different pixels pair distances.

TABLE II
EXAMPLES AND THE NUMBER OF QUERIES OF EACH CATEGORY IN OUR DATASET

Category	Number of queries	Examples
Animal	100	*Alligator*, *Bat*, *Cattle*
Cartoon	92	*Air gear*, *Final fantasy*
Event	78	*Olympic*, *Wedding*, *WWE*
Object	295	*Airplane*, *Bed*, *Toy*
People	68	*Girls*, *Snowman*, *Baby*
Person	40	*Tom Hanks*, *Will Smith*
Scene	48	*Dersert*, *Rainbow*
Time08	88	*Barack Obama*, *Steve Jobs*
Misc	288	*Japan*, *Titanic*, *Adidas*

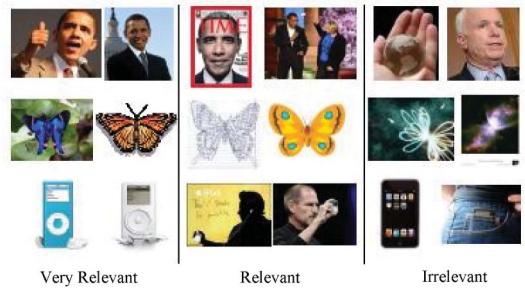


Fig. 3. Several example images with different relevance levels to *Barack Obama, Butterfly*, and *ipod*, respectively.

4) 256-dimensional RGB color histogram. A 256-dimensional histogram feature vector is extracted in RGB color space.
5) 75-dimensional Edge distribution histogram. Each image is divided into 5 blocks and 15-dimensional EDH features are extracted.
6) 128-dimensional Wavelet texture. 128-dimensional features are extracted using the mean and standard deviation of the energy distribution of each sub-band at different levels.
7) 7-dimensional Face features. The features include the number of faces, the ratio of face areas and the position of the largest face region.

More details about the features can be found in [45].

We adopt NDCG [46] as the performance evaluation measure. The NDCG measure is computed as

$$NDCG@P = Z_P \sum_{i=1}^{P} \frac{2^{l(i)} - 1}{\log(i+1)} \tag{25}$$

where P is the considered depth, $l(i)$ is the relevance level of the i-th image and Z_P is a normalization constant that is chosen to let the optimal ranking's NDCG score to be 1.

B. On the Integration of Multiple Modalities

In this part, we compare our approach that integrates all modalities with the methods that use only an individual modality. We denote the proposed method as "MGL" (multimodal graph-based learning) and use "MGL-CM", "MGL-HSV", "MGL-CORR", "MGL-RGB", "MGL-EDH", "MGL-Wavelet" and "MGL-Face" to denote the methods that only use the seven modalities, respectively.

For the proposed "MGL" method, there are two parameters, i.e., λ and ξ (see Eq. (11)). We tune the two parameters based on an additional small dataset. Specifically, we jointly tune the two parameters to optimize the reranking performance of the "MGL" method on the MSRA-MM Version 1.0 dataset [45], which contains 68 queries. Since there is no overlap between the MSRA-MM Version 1.0 and 2.0 queries, there will be no over-fitting effect. For the other seven methods, there is only one parameter, i.e., λ. We also tune the parameter based on the MSRA-MM Version 1.0 dataset. The neighborhood size N is set to 20. For the iteration times T, T_1 and T_2, we set them to 5, 10 and 10, respectively (in our experiments we found that these values can lead to a well convergence of the alternating optimization process).

Table III illustrates the average NDCG@100 measurements obtained by different methods for each category of queries. Here we have also illustrated the NDCG@100 measurements of the original ranking lists and we regard them as "Baseline" results. From the average results we can see that nearly all the methods can effectively improve the baseline results, except "MGL-Face". This is because the 7-dimensional face-relative features are not informative enough and thus the reranking introduces performance degradation. But they are still useful by integrating them with other features to work together. The "MGL" approach that integrates all the modalities achieves the best results. While all the average NDCG@100 measurements are all below 0.8, it can achieve a measurement of 0.816.

To further analyze the results, we consider the improvement levels brought by each reranking method. Table IV illustrates the distribution of the relative improvements brought by each reranking method. From the results we can see that several queries can get encouraging improvements while several others have degraded performance. This is a well-known phenomenon, i.e., reranking will not always help in improving performance [8], [47]. The "MGL" approach demonstrates the most robust performance. For 82.7% of the queries, the "MGL" approach can improve the original ranking lists. This compares favorably with the other methods that use an individual modality. For 23.7% of the queries, the relative improvements are above 20%.

C. On the Comparison of Different Reranking Approaches

We then compare the proposed "MGL" approach with several existing reranking approaches, including:

1) Baseline, i.e., the original search results without reranking.
2) The information bottleneck based clustering method in [14]. The reranking approach adopts a pseudo-relevance feedback and information bottleneck clustering over visual features with the help of a smoothed initial ranking. The method is denoted as "Clustering".
3) The random walk method proposed in [23], which estimates the relevance scores of images by performing a random walk. The method is denoted as "Random Walk".

TABLE III

COMPARISON OF THE AVERAGE NDCG@100 MEASUREMENTS OBTAINED BY INTEGRATING ALL MODALITIES AND USING ONLY AN INDIVIDUAL MODALITY. FROM THE RESULTS WE CAN SEE THAT THE MGL, WHICH INTEGRATES MULTIPLE MODALITIES, OUTPERFORMS THE OTHER METHODS THAT USE AN INDIVIDUAL MODALITY

Category \ Method	Baseline	MGL-CM	MGL-HSV	MGL-CORR	MGL-RGB	MGL-EDH	MGL-Wavelet	MGL-Face	MGL
Animal	0.734	0.746	0.741	0.765	0.729	0.742	0.750	0.743	**0.791**
Cartoon	0.807	0.827	0.829	0.849	0.825	0.837	0.831	0.820	**0.865**
Event	0.788	0.785	0.788	0.806	0.787	0.789	0.8	0.783	**0.811**
Object	0.703	0.719	0.705	0.718	0.696	0.692	0.717	0.708	**0.745**
People	0.714	0.711	0.706	0.728	0.696	0.708	0.724	0.669	**0.742**
Person	0.908	0.920	0.924	0.927	0.917	0.924	0.924	0.926	**0.940**
Scene	0.703	0.742	0.736	0.758	0.732	0.716	0.744	0.715	**0.792**
Time08	0.830	0.846	0.835	0.855	0.829	0.833	0.825	0.608	**0.870**
Misc	0.736	0.760	0.754	0.764	0.747	0.756	0.757	0.725	**0.790**
Mean	0.769	0.784	0.780	0.797	0.773	0.777	0.786	0.744	**0.816**

TABLE IV

DISTRIBUTION OF RELATIVE PERFORMANCE IMPROVEMENTS BY EACH RERANKING METHOD AMONG THE 1096 QUERIES

Improvement \ Method	MGL-CM	MGL-HSV	MGL-CORR	MGL-RGB	MGL-EDH	MGL-Wavelet	MGL-Face	MGL
Below -20%	0.121	0.134	0.123	0.161	0.158	0.110	0.054	0.070
-20% to -10%	0.105	0.098	0.082	0.107	0.087	0.095	0.058	0.048
-10% to -5%	0.068	0.086	0.075	0.073	0.083	0.069	0.058	0.024
-5% to 0	0.079	0.093	0.091	0.088	0.079	0.100	0.072	0.030
0 to 5%	0.276	0.231	0.264	0.234	0.256	0.275	0.639	0.186
5% to 10%	0.101	0.109	0.085	0.083	0.094	0.091	0.069	0.163
10% to 20%	0.110	0.102	0.130	0.126	0.115	0.119	0.031	0.204
Above 20%	0.141	0.148	0.151	0.130	0.130	0.142	0.018	0.274

4) Bayesian reranking proposed in [17]. We concatenate all features into a long vector and then perform the strength based method in [17]. The method is denoted as "Bayesian".
5) Graph-based reranking with concatenated features. That is, we concatenate all the features into a long vector and then perform the graph-based reranking shown in Eq. (11) by setting K to 1. The method is denoted as "Concatenated Features".
6) Pseudo relevance feedback. Given a query, we use the top 100 search results in the original ranking list as positive samples, and then randomly collect 100 images from the whole database and regard them as negative samples. We then learn a support vector machine classifier with RBF kernel based on these samples and use the classifier to rerank the search results. The method is denoted as "PRF".
7) Late fusion with tuned weights. That means, we fuse the results of "MGL-CM", "MGL-HSV", "MGL-CORR", "MGL-RGB", "MGL-EDH", "MGL-Wavelet" and "MGL-Face". The weights are tuned to their optimal values based on the MSRA-MM Version 1.0 queries. The method is denoted as "Late Fusion".
8) Multimodal graph-based reranking with assigning equivalent weights to all modalities. That means we fix $\alpha_i = 1/K$. The method is denoted as "MGL (Equal Weights)".
9) Multimodal graph-based reranking with heuristic weights assigned to different modalities. We assign the weights that are proportional to the performance gains in Table III, such that more effective modalities can get higher weights. The method is denoted as "MGL (Heuristic Weights)".

Each of the above methods involves several parameters. We tune all these parameters to their optimal values on the MSRA-MM version 1.0 dataset, which is similar to the process introduced in the above subsection. In this way, we can provide a fair comparison for these algorithms.

Figure 4 demonstrates the top results obtained by different methods for an example query *amber*. Table V illustrates the average NDCG@100 measurements obtained by different methods for each category. From the results we can see that the proposed "MGL" approach shows the best average performance for each type of queries. This demonstrates the robustness of this algorithm. In particular, we can see that the "Concatenated Features" only achieves very limited performance improvement over the "Baseline" results. This demonstrates that, although we have a distance metric learning component that can somewhat

Fig. 4. Top results in the original ranking list (baseline) and the reranked lists obtained by different methods for an example query *amber*. The proposed MGL method obtains the best results; the top images are all relevant. The other methods contain at least one irrelevant image in the top results. (a) Baseline. (b) Clustering. (c) Random walk. (d) Bayesian. (e) Concatenated features. (f) PRF. (g) Late fusion. (h) MGL (equal weights). (i) MGL (heuristic weights). (j) MGL.

modulate the effects of different features, the graph-based reranking still cannot well handle the high-dimensional features. The "Late Fusion", "MGL (Equal Weights)" and "MGL (Heuristic Weights)" methods cannot achieve sufficiently good performance because they are unable to adaptively modulate the effects of multiple modalities for different queries. Although in the "MGL (Heuristic Weights)" method we have set weights according to the performance gains of different modalities in Table III, it is still not as reasonable as the proposed approach as the description ability of a modality should vary across queries. The "PRF" method is the fastest as it only needs to train a classification model with several pseudo positive and negative examples, but we can see that it performs much worse than the proposed "MGL" method. The "Bayesian" method performs the second best among all the compared approaches. This can be partially attributed to the fact that the "Bayesian" method adopts a more reasonable loss term. The loss term is built based on a preference strength and it is better than the squared distance of relevance scores (see [17]). We also compare the NDCG measures with different depths of these methods. Figure 5 demonstrates the average NDCG@3, NDCG@10, NDCG@20, NDCG@50 and NDCG@100 measurements obtained by these methods, and we can see that

TABLE V

COMPARISON OF THE AVERAGE NDCG@100 MEASUREMENTS OBTAINED BY DIFFERENT METHODS FOR EACH CATEGORY OF QUERIES

Category \ Method	Baseline	Clustering	Random Walk	Bayesian	Concatenated Features	PRF	Late Fusion	MGL (Equal weights)	MGL (Heuristic weights)	MGL
Animal	0.734	0.759	0.753	0.773	0.724	0.750	0.758	0.768	0.773	**0.791**
Cartoon	0.807	0.828	0.827	0.844	0.819	0.829	0.831	0.83	0.84	**0.865**
Event	0.788	0.788	0.808	0.803	0.779	0.775	0.795	0.787	0.789	**0.811**
Object	0.703	0.717	0.721	0.731	0.708	0.708	0.715	0.728	0.733	**0.745**
People	0.714	0.710	0.732	0.724	0.716	0.715	0.715	0.717	0.724	**0.742**
Person	0.908	0.905	0.931	0.922	0.939	0.913	0.91	0.922	0.923	**0.940**
Scene	0.703	0.752	0.721	0.766	0.712	0.761	0.75	0.758	0.766	**0.792**
Time08	0.830	0.854	0.851	0.870	0.830	0.860	0.852	0.858	0.863	**0.870**
Misc	0.736	0.753	0.755	0.767	0.757	0.758	0.758	0.757	0.76	**0.790**
Mean	0.769	0.785	0.789	0.800	0.776	0.785	0.787	0.792	0.797	**0.816**

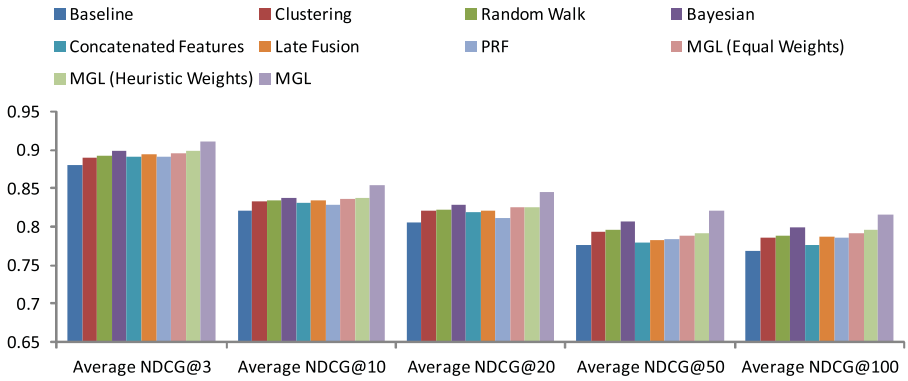


Fig. 5. NDCG measurements with different depths obtained by the compared reranking methods.

the proposed "MGL" approach consistently achieves the best performance.

We further perform a statistical significance test to verify whether the superiority of the "MGL" method is statistically significant. The p values of the t-test of the "MGL" method over the other methods, including those that use only an individual modality, are shown in Table VI. From the results we can see that the superiority of the "MGL" method is statistically significant.

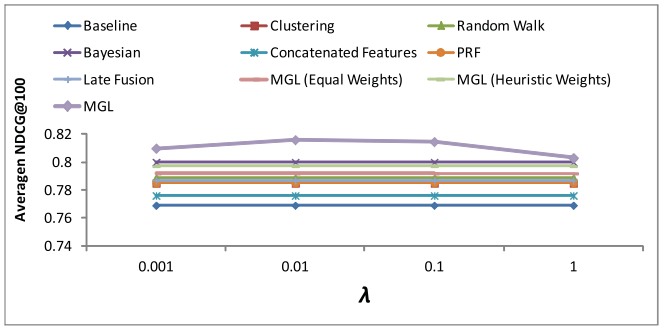
TABLE VI

p VALUES OF THE SIGNIFICANCE TEST

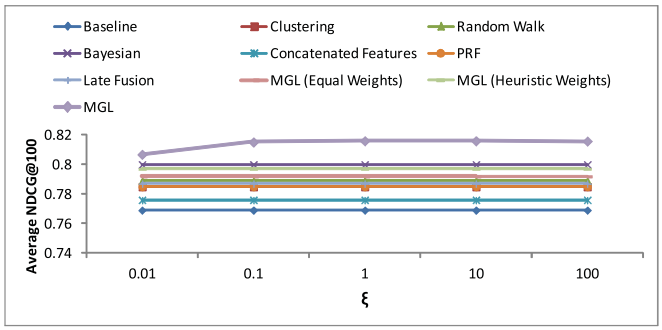
Comparison	p
MGL versus MGL-CM	$4.35 \times e^{-10}$
MGL versus MGL-HSV	$5.75 \times e^{-8}$
MGL versus MGL-CORR	0.002
MGL versus MGL-RGB	$1.11 \times e^{-16}$
MGL versus MGL-EDH	$2.62 \times e^{-10}$
MGL versus MGL-Wavelet	$1.59 \times e^{-7}$
MGL versus MGL-Face	$1.01 \times e^{-16}$
MGL versus Baseline	$1.2 \times e^{-12}$
MGL versus Clustering	$1.49 \times e^{-7}$
MGL versus Random Walk	$2.54 \times e^{-10}$
MGL versus Bayesian	$1.08 \times e^{-5}$
MGL versus Concatenated Features	$1.73 \times e^{-11}$
MGL versus PRF	$7.34 \times e^{-8}$
MGL versus Late Fusion	$7.39 \times e^{-9}$
MGL versus MGL (Equal Weights)	$1.06 \times e^{-6}$
MGL versus MGL (Heuristic Weights)	$5.38 \times e^{-6}$

D. On the Parameters λ and ξ

Finally, we also test the sensitivity of the two parameters λ and ξ, which are used in the proposed algorithm. We first set ξ to 1 and vary λ from 0.001 to 1. Figure 6(a) demonstrates the performance curve with respect to the variation of λ. We then set λ to 0.01 and vary ξ from 0.01 to 100. Figure 6(b) demonstrates the performance curve with respect to the variation of ξ. Here we also illustrate the performance of the other nine methods, i.e., "Baseline", "Clustering", "Random Walk", "Bayesian", "Concatenated Features", "PRF", "Late Fusion", "MGL (Equal Weights)" and "MGL (Heuristic Weights)", for comparison. From the results we can see that the performance of our approach will not significantly degrade when the two parameters vary in a fairly wide range and it can keep outperforming the other nine methods.



(a)



(b)

Fig. 6. Illustration of the effects of the parameters λ and ξ in reranking. (a) Reranking performance variation when λ varies from 0.001 to 1. (b) Reranking performance variation when ξ varies from 0.01 to 100.

V. Conclusion

This paper introduces a web image search reranking approach that explores multiple modalities in a graph-based learning scheme. The approach simultaneously learns relevance scores, weights of modalities, and the distance metric and its scaling for each modality. To test the performance of the proposed approach, we have conducted experiments on a dataset that contains 1,096 queries. The effectiveness of integrating multiple modalities has been demonstrated. It is demonstrated that the proposed approach not only achieves better average results but also shows more robustness than the methods that use only an individual modality. We have also compared our approach with several existing reranking methods, and results also demonstrate the superiority of our approach.

We only consider search relevance in this work, but actually diversity is also an important aspect for search performance. In fact, after performing reranking, we can further have a diversification process to enhance the diversity of top search results, such as by using the method proposed in [48].

References

[1] H. Li, M. Wang, Z. Li, Z. J. Zha, and J. Shen, "Optimizing multimodal reranking for web image search," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 1119–1120.

[2] M. Wang, X. S. Hua, J. Tang, and R. Hong, "Beyond distance measurement: Constructing neighborhood similarity for video annotation," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 465–476, Apr. 2009.

[3] M. Wang, X. S. Hua, R. Hong, J. Tang, G. J. Qi, and Y. Song, "Semi-supervised kernel density estimation for video annotation," *Comput. Vis. Image Understand.*, vol. 113, no. 3, pp. 384–396, 2009.

[4] J. Yu, D. Tao, and M. Wang, "Adaptive hypergraph learning and its application in image classification," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3262–3272, Jul. 2012.

[5] L. Duan, W. Li, I. W. Tsang, and D. Xu, "Improving web image search by bag-based reranking," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3280–3290, Nov. 2011.

[6] V. Jain and M. Varma, "Learning to re-rank: Query-dependent image re-ranking using click data," in *Proc. 20th Int. World Wide Web Conf.*, 2011, pp. 277–286.

[7] J. Krapac, M. Allan, J. Verbeek, and F. Jurie, "Improving web image search results using query-relative classifiers," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1094–1101.

[8] N. Morioka and J. Wang, "Robust visual reranking via sparsity and ranking constraints," in *Proc. ACM Multimedia*, 2011, pp. 533–542.

[9] X. Tian and D. Tao, "Visual reranking: From objectives to strategies," *IEEE Multimedia*, vol. 18, no. 3, pp. 12–21, Mar. 2011.

[10] H. Zitouni, S. Sevil, D. Ozkan, and P. Duygulu, "Re-ranking of web image search results using a graph algorithm," in *Proc. Int. Conf. Image Process.*, 2008, pp. 1–4.

[11] X. Tian, D. Tao, X. S. Hua, and X. Wu, "Active reranking for web image search," *IEEE Trans. Image Process.*, vol. 19, no. 3, pp. 805–820, Mar. 2010.

[12] X. He, W. Y. Ma, and H. J. Zhang, "Learning an image manifold for retrieval," in *Proc. ACM Multimedia*, 2004, pp. 17–23.

[13] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[14] W. Hsu, L. S. Kennedy, and S. F. Chang, "Video search reranking through random walk over document-level context graph," in *Proc. ACM Multimedia*, 2007, pp. 971–980.

[15] Y. Liu, T. Mei, and X. S. Hua, "CrowdReranking: Exploring multiple search engines for visual search reranking," in *Proc. ACM SIGIR*, 2009, pp. 500–507.

[16] Y. Liu, T. Mei, X. Q. Wu, and X. S. Hua, "Multigraph-based query-independent learning for video search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 12, pp. 1841–1850, Dec. 2009.

[17] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X. S. Hua, "Bayesian video search reranking," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 131–140.

[18] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. ACM Multimedia*, 2005, pp. 399–402.

[19] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proc. Int. Conf. Image Video Retrieval*, 2003, pp. 238–247.

[20] A. P. Natsev, M. R. Naphade, and J. Tesic, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *Proc. ACM Multimedia*, 2005, pp. 598–607.

[21] Y. Liu and T. Mei, "Optimizing visual search reranking via pairwise learning," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 280–291, Apr. 2011.

[22] L. Kennedy and S. F. Chang, "A reranking approach for context-based concept fusion in video indexing and retrieval," in *Proc. 6th ACM Int. Conf. Image Video Retrieval*, 2007, pp. 333–340.

[23] Y. Jing and S. Baluja, "VisualRank: Applying pagerank to large-scale image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1877–1890, Nov. 2008.

[24] L. Yang and A. Hanjalic, "Supervised reranking for web image search," in *Proc. ACM Multimedia*, 2010, pp. 183–192.

[25] T. Yao, T. Mei, and C. W. Ngo, "Co-reranking by mutual reinforcement for image search," in *Proc. Int. Conf. Image Video Retrieval*, 2010, pp. 34–41.

[26] S. Clinchant, J. M. Renders, and G. Csurka, "Trans-media pseudo-relevance feedback methods in multimedia retrieval," in *Proc. Image Retrieval CLEF*, 2007, pp. 569–576.

[27] D. Kilinc and A. Alpkocak, "DEU at imageCLEF 2009 wikipediaMM task: Experiments with expansion and reranking approaches," in *Proc. Image Retrieval CLEF*, 2009, pp. 1–9.

[28] K. W. Wan, Y. T. Zheng, and S. Roy, "I2R at imageCLEF wikipedia retrieval 2010," in *Proc. Image Retrieval CLEF*, 2010, pp. 1–9.

[29] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proc. ACM Multimedia*, 2004, pp. 572–579.

[30] G. Iyengar, H. J. Nock, and C. Neti, "Discriminative model fusion for semantic concept detection and annotation in video," in *Proc. ACM Multimedia*, 2003, pp. 255–258.

[31] R. Yan and A. Hauptmann, "The combination limit in multimedia retrieval," in *Proc. ACM Multimedia*, 2003, pp. 339–342.

[32] M. Wang, X. S. Hua, R. Hong, J. Tang, G. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, May 2009.
[33] B. Geng, C. Xu, D. Tao, L. Yang, and X. S. Hua, "Ensemble manifold regularization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2396–2402.
[34] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 1–8.
[35] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 912–919.
[36] J. He, M. Li, H. J. Zhang, H. Tong, and C. Zhang, "Manifold-ranking based image retrieval," in *Proc. ACM Multimedia*, 2004, pp. 9–16.
[37] X. Yuan, X. S. Hua, M. Wang, and X. Wu, "Manifold-ranking based video concept detection on large database and feature pool," in *Proc. ACM Multimedia*, 2006, pp. 623–626.
[38] Y. Gao, M. Wang, Z. J. Zha, Q. Tian, Q. Dai, and N. Zhang, "Less is more: Efficient 3-D object retrieval with query view selection," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1007–1018, Oct. 2011.
[39] B. Geng, D. Tao, and C. Xu, "DAML: Domain adaptation metric learning," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2980–2989, Oct. 2011.
[40] N. Sebe, M. S. Lew, and D. P. Huijsmans, "Toward improved ranking metrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1132–1143, Oct. 2000.
[41] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 929–942, Jul. 2010.
[42] F. Wang and C. Zhang, "Label propagation through linear neighborhoods," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 55–67, Jan. 2008.
[43] X. Zhang and W. S. Lee, "Hyperparameter learning for graph-based semi-supervised learning algorithms," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2006, pp. 1–8.
[44] J. He, M. Li, H. J. Zhang, H. Tong, and C. Zhang, "Generalized manifold-ranking-based image retrieval," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3170–3177, Oct. 2006.
[45] H. Li, M. Wang, and X. S. Hua, "MSRA-MM 2.0: A large-scale web multimedia dataset," in *Proc. Int. Conf. Data Mining Workshop*, 2009, pp. 164–169.
[46] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.
[47] X. Tian, Y. Lu, L. Yang, and Q. Tian, "Learning to judge image search results," in *Proc. 19th ACM Multimedia*, 2011, pp. 363–372.
[48] M. Wang, K. Yang, X. S. Hua, and H. J. Zhang, "Toward a relevant and diverse search of social images," *IEEE Trans. Multimedia*, vol. 12, no. 8, pp. 829–842, Dec. 2010.

Meng Wang (M'09) received the B.E. degree from the Special Class for the Gifted Young, and the Ph.D. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively.

He is currently a Professor with the Hefei University of Technology, Hefei. He was an Associate Researcher with Microsoft Research Asia, Shanghai, China, a Core Member with a startup in Silicon Valley, and a Senior Research Fellow with the National University of Singapore, Singapore. He has authored or co-authored more than 100 book chapters, and journal and conference papers in his areas of expertise. His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing.

Dr. Wang was a recipient of the Best Paper Award in the 17th and 18th ACM International Conference on Multimedia and the Best Paper Award in the 16th International Multimedia Modeling Conference. He is a member of ACM.

Hao Li received the B.E. degree in software engineering from Shandong University, Jinan City, China, in 2009 and the M.S. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently pursuing the Ph.D. degree with the University of Maryland, College Park.

He was a Research Intern with the Internet Graphics Group and Media Computing Group, Microsoft Research Asia, Shanghai, China, in 2010 and 2009, respectively. His current research interests include computer vision, information retrieval, and multimedia analysis.

Dacheng Tao (M'07–SM'12) is currently a Professor of computer science with the Centre for Quantum Computation and Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia. His current research include statistics and mathematics for data analysis problems in computer vision, machine learning, multimedia, data mining, and video surveillance. He has authored or co-authored more than 100 scientific articles at top venues, including the IEEE T-PAMI, T-IP, T-SP, CVPR, ECCV, AISTATS, ICDM, ACM, and Multimedia and SIGKDD.

Dr. Tao was a recipient of the Best Theory and Algorithm Paper Runner Up Award at the Seventh IEEE Conference on Data Mining in 2007 (ICDM'07).

Ke Lu received the Masters degree from the Department of Mathematics and the Ph.D. degree from the Department of Computer Science, Northwest University, Evanston, IL, in 1998 and 2003, respectively.

He is currently a Professor with the Graduate University of Chinese Academy of Sciences, Beijing, China. His current research interests include curve matching, 3-D image reconstruction, and computer graphics.

Xindong Wu (F'10) received the Bachelor's and Master's degrees in computer science from the Hefei University of Technology, Hefei, China, in 1984 and 1987, respectively, and the Ph.D. degree in artificial intelligence from the University of Edinburgh, Edinburgh, U.K. in 1993.

He is currently a Yangtze River Scholar with the School of Computer Science and Information Engineering, Hefei University of Technology, and a Professor of computer science with the University of Vermont, Burlington. His current research interests include data mining, knowledge-based systems, and Web information exploration.

Dr. Wu is the Steering Committee Chair of the IEEE International Conference on Data Mining (ICDM), the Editor-in-Chief of Knowledge and Information Systems (KAIS, by Springer), and a Series Editor of the Springer Book Series on Advanced Information and Knowledge Processing. He was the Editor-in-Chief of the IEEE Transactions on Knowledge and Data Engineering (TKDE, by the IEEE Computer Society) from 2005 to 2008. He served as the Program Committee Chair or the Co-Chair for ICDM'03, KDD-07 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), and CIKM 2010 (the 19th ACM Conference on Information and Knowledge Management).