2-2011

A Two-View Learning Approach for Image Tag Ranking

Jinfeng ZHUANG
Nanyang Technological University

Steven C. H. HOI
Singapore Management University, CHHOI@smu.edu.sg

DOI: https://doi.org/10.1145/1935826.1935913

A Two-View Learning Approach for Image Tag Ranking

Jinfeng Zhuang
School of Computer Engineering
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
zhua0016@ntu.edu.sg

Steven C.H. Hoi
School of Computer Engineering
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
chhoi@ntu.edu.sg

ABSTRACT

Tags of social images play a central role for text-based social image retrieval and browsing tasks. However, the original tags annotated by web users could be noisy, irrelevant, and often incomplete for describing the image contents, which may severely deteriorate the performance of text-based image retrieval models. In this paper, we aim to overcome the challenge of *social tag ranking* for a corpus of social images with rich user-generated tags by proposing a novel two-view learning approach. It can effectively exploit both textual and visual contents of social images to discover the complicated relationship between tags and images. Unlike the conventional learning approaches that usually assume some parametric models, our method is completely data-driven and makes no assumption of the underlying models, making the proposed solution practically more effective. We formally formulate our method as an optimization task and present an efficient algorithm to solve it. To evaluate the efficacy of our method, we conducted an extensive set of experiments by applying our technique to both text-based social image retrieval and automatic image annotation tasks, in which encouraging results showed that the proposed method is more effective than the conventional approaches.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

Keywords

Tag ranking, social images, image search, optimization, annotation, two-view learning, recommendation

1. INTRODUCTION

In the web 2.0 era, along with the popularity of various digital imaging devices and the advances of Internet technologies, digital images and photos can be easily created, uploaded, shared and distributed on the World Wide Web (WWW). Unlike the situation one decade ago, web images are nowadays playing a more and more important role in the WWW today. Web image search thus has become an active yet rather challenging research issue.

Figure 1: An illustration of the efficacy of tag ranking. The original annotations contain meaningless tags, and the most relevant tag "panda" is ranked at the fifth position. After tag ranking, "panda" comes up to the top position. The noisy tags are removed.

One major difficulty for web image search is that most images are usually not annotated with proper tags, and many of them are even completely unlabeled. In addition, even for the annotated images, their associated tags could be noisy, irrelevant, and often incomplete for describing the contents of the images. According to the study in [12] that conducted the statistics on Flickr [1], a popular web 2.0 image sharing portal, only 50% of tags are actually related to the image content. This poses a great challenge for typical web image search approaches based on existing web search engine solutions, which often simply apply regular text based retrieval techniques on the web image search domain.

To address this challenge, one approach is to study the tag refinement techniques, which have been proposed by many researchers recently [9, 20, 21, 11, 25, 12]. Although tag refinement techniques can generally improve the quality of the tags, they do not explicitly answer which tag is more relevant than the other for a specific image. Very recently, Liu et. al. [12] proposed the *tag ranking* problem as an alternative solution to address the social tag issue. In general, the goal of social tag ranking is to rank the tags of a social image according to their relevance to the semantic/visual content of the image. Figure 1 illustrates an example of showing

[1] http://www.flickr.com/

WSDM'11, February 9–12, 2011, Hong Kong, China.

the lists of tags before and after the process of social tag ranking, in which the relevant tags can be ranked in the top positions using the proposed social tag ranking technique. Social tag ranking is important as it can facilitate a lot of real-world multimedia applications, including social image retrieval, browsing, and annotation tasks.

Despite the encouraging results reported by the study in [12], there remains some limitations for the existing work. First of all, the existing work usually assume certain parametric functions to model the tag generation process. The probabilistic approaches limit the capability of fitting the complicated image and tag relationship as the parametric probabilistic model assumption seldom holds in practice. In addition, the existing studies often adopt heuristic tag ranking methods without careful optimization, which can hardly achieve the optimal results.

Unlike the existing work, this paper proposes a novel two-view learning approach for social tag ranking, which is purely data-driven, i.e., without assuming explicit relevance models between tags and images. Specifically, we formulate the tag ranking task as a problem of learning a *nonparametric* tag weighting matrix that encodes the relevance relationship between images and tags. We then present an effective algorithm to optimize the weight matrix by exploiting both local visual geometry in image space and local textual geometry in tag space.

In sum, the major contributions of our work include:

- We propose a novel two-view learning framework for social tag ranking, which is purely data-driven without making any assumption on modeling the relationship between images and tags. Thus it is more flexible and powerful to learn the complicated relationship in real-world social image data.
- We formulate the two-view tag weighting problem as an optimization task, and present an effective stochastic coordinate descent algorithm, which can solve the optimization efficiently.
- We conduct an extensive set of experiments to examine the empirical performance of the proposed tag ranking technique and apply our technique for several applications, including social image retrieval and tag recommendation for web photo tagging.

The rest of this paper is organized as follows. Section 2 discusses some related work. Section 3 presents the proposed two-view learning approach for social image tag ranking. Section 4 conducts an extensive set of experiments to evaluate the proposed methods. Section 5 discusses some limitations of our study, and Section 6 concludes this work.

2. RELATED WORK

The quality of tags play a crucial role in social image retrieval. Recent years have witnessed a lot of emerging studies to address the tag quality issues. In this section, we summarize and analyze some representative methods that are closely related to the techniques presented in this paper.

The first category of related techniques refers to *tag annotation*. Annotating an image automatically by machine learning methods enables large amounts of unlabeled images to be indexed and searchable by existing text based image search engines. A variety of techniques have been proposed for auto-image annotation in recent years [7, 4, 15, 22, 13, 17, 24]. In general, auto-image annotation can be viewed as an intermediate task for a generic web image retrieval task. Most of these methods try to model the probabilistic relationship between tags and images in one way or another. However, producing highly accurate annotation results remains an unsolved long-term challenge.

Instead of auto-annotation, an alternative approach is to study *tag refinement* [9, 20, 21, 5], which aims to model the relevance of the associated tags to an image. Jin et. al. [9] proposed the pioneering work on annotation refinement by a generic data-based WordNet. The assumption is that highly correlated annotation tends to be correct and non-correlated tags tend to be noisy. With a large collection of social images, one could build the correlation matrix among the tags. However, this method simply ignores the specific content information of each individual image.

To address the limitation of WordNet, Wang et. al. [20] proposed the Random Walk with Restarts (RWR) algorithm. The key idea is to not only use the co-occurrence-based tag similarity, but also leverage the information of the original annotated order of tags. Further, Wang et. al. [21] proposed the Content-based Image Annotation Refinement (CIAR) algorithm that formulates the tag refinement problem as a Markov process and the candidate tags are defined as the states of a Markov chain. The transition matrix is constructed based on the query image using both visual features of the query image and the corpus information. The CIAR algorithm focuses on refining the automatic annotation results of a query image. If the goal is to refine the existing tags of a large corpus, the computation of query-based transition matrix could be potentially highly intensive. Weinberger et. al. [23] proposed a probabilistic approach to modeling the ambiguity of tags. Different from the previous methods, it can suggest tags that are not included in the user-generated tag list. Li et. al. [11] proposed a voting method by first finding the nearest neighbors of the given image, and collect the votes from the nearest neighbors. The tag relevance is determined based on the number of such votes from the nearest neighbors.

Very recently, researchers are interested in a specific tag refinement task, known as "tag ranking" [12], which aims to generate a permutation of the associated tags for an image, in which the resulting order indicates the tags' relevance or importance to the image. Although some existing tag refinement methods might be somehow adapted to the tag ranking task, the *tag ranking* problem is explicitly addressed very recently by the study in [12]. In their algorithm, they first model the generating probability of tags from an image with some exponential function. After that they refine the ranking score by random walk over a similarity matrix between tags constructed by incorporating both the representative image of that tag and the Google distance between pairs of tags. Despite encouraging results reported, their method assumes some parametric models that may limit their capability of modeling complicated tag-image relationship.

The importance of tag ranking calls for further study on this problem. In this paper, we propose a novel two-view learning approach without assuming any parametric models between images and tags, which distinguishes our work from the existing model-based methods. Finally, we formally formulate our method as an optimization task, which differs from the other heuristic tag ranking methods.

3. TWO-VIEW LEARNING FOR SOCIAL TAG RANKING

In this section we present a novel two-view learning approach for social tag ranking. We first give some preliminaries to introduce our problem setting and present the two-view representation for modeling social images. We then present the proposed learning framework and an efficient algorithm to solve the optimization followed by the discussion on some practical implementation issues.

3.1 Preliminaries

Consider a social image z is represented as a pairwise example (x, t), which consists of an image $x \in \mathcal{X}$ and its associated set of tags $t \subseteq \mathcal{T}$, where $\mathcal{X}$ and $\mathcal{T}$ are referred to the image and tag spaces, respectively. In the sequel, we let n denote the number of social images in the corpus, and $m = |\mathcal{T}|$ denote the number of unique tags in the corpus. For a positive integer d, we define $\mathbb{N}_d = \{1, \ldots, d\}$ as a series of d . For any matrix M, we use the following notation:

- M_i denotes the i-th row vector of M;
- M_i^c denotes the i-th column vector of M;
- M_{ij} denotes the (i, j)-th entry of M;
- $M^\top$ denotes the transpose of M;
- $\|M\|_F = \sqrt{\sum_{ij} M_{ij}^2}$ is the Frobenius norm of M;
- tr $M = \sum_i M_{ii}$ is the trace of M if M is square.

For every social image $z_i = (x_i, t_i)$, the set of associated tags is assigned to z_i provided by web users. Without further information, we treat all the assigned tags equally important. Therefore, we indicate the initial annotation of z_i by a vector $t_i \in \mathbb{R}^m$, where $t_{ij} = 1/|t_i|$ if the j-th tag is assigned to z_i; $t_{ij} = 0$ otherwise. We tile all t_i values into a tag indicator matrix T such that $T_{ij} := t_{ij}$.

Based on the above definitions, for a given social image $z_i = (x_i, t_i)$, social tag ranking in general is to find an optimal permutation of the tag list t_i by learning from a corpus of social images $\{z_i = (x_i, t_i), i \in \mathbb{N}_n\}$. It is however hard to directly optimize the permutation of the tag list t_i. In this paper, we consider an alternative approach by learning a nonparametric tag weighing matrix $W \in \mathbb{R}_+^{n \times m}$, where entry $w_{ij}, i \in \mathbb{N}_n, j \in \mathbb{N}_m$, indicates the relevance of tag t_{ij} with respect to image x_i.

As a result, the problem of *social tag ranking* is equivalent to looking for the optimal tag weighting matrix from mining the hidden knowledge from the social image corpus $\{z_i = (x_i, t_i), i \in \mathbb{N}_n\}$. Below we propose a purely data-driven approach to learning the optimal tag weighting matrix W. Our method does not make explicit parametric assumptions on any generative models between $\mathcal{T}$ and $\mathcal{X}$.

3.2 Two-View Representation for Social Tags

The social image data Z consists of two views: the visual view ($\mathcal{X}$) and the tag or concept view ($\mathcal{T}$). As mentioned above, to model the relationship between images and tags, we employ a nonparametric tag weighting matrix W that has a natural interpretation for the two-view representation of social images:

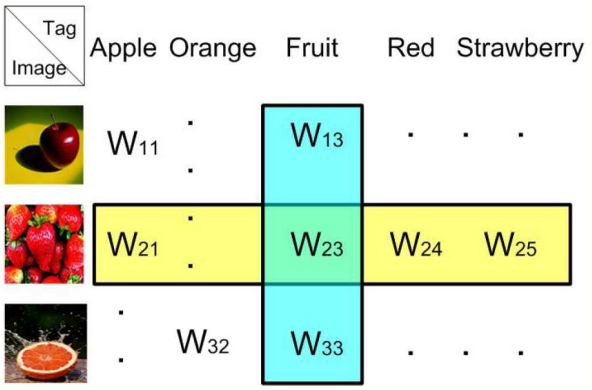

Figure 2: Illustration of the two views of the weight matrix of social images. Each W_{ij} is the weight of the j-th tag for the i-th image. The weights in the vertical blue ellipse provides an exempler representation of "fruit". The weights in the horizontal yellow ellipse provides a semantic representation of the image strawberry.

- *Row-Pivot* view: Each row vector of W, denoted as W_i, is a weighting vector over a set of user-generated tags, which forms a semantic summarization for describing the semantic content of the visual image x_i. From this point of view, we can represent a visual image as a weighted combination of a set of relevant tags via specifying the row of W in tag space.
- *Column-Pivot* view: Each column vector of W, denoted as W_i^c, is a weighting vector over a set of images, which actually forms an exemplar of the corresponding tag. From this point of view, we can represent a tag / a semantic concept as a weighted combination of a set of most representative images via specifying the column of W in visual space.

To better understand the idea, we give a visual example to illustrate the two-view representation as shown in Figure 2. For the highlighted row and column in this Figure, the weights in the vertical (blue) zone provides an exampler representation of the tag "fruit", and the weights in the horizontal (yellow) zone provides a semantic representation of the image "strawberry". From the above observation, we can see that the relevance weighing matrix W plays a central role for modeling the relationship between $\mathcal{X}$ and $\mathcal{T}$.

3.3 Two-View Learning for Tag Weighting

The basic idea of learning the optimal W is twofold: (1) we aim to make the above two-view representations coincide with the local geometry in both visual space and concept space; (2) we shall preserve the user annotation results to some extent. Motivated by these two considerations, we propose to devise the following learning scheme.

For the first purpose, we have to make use of the similarity graph S^x in $\mathcal{X}$ and S^t in $\mathcal{T}$ (we assume both S^x and S^t are symmetric). We take S^x to sketch the details. For any two images x_i and x_j, the entry S_{ij}^x computes the visual similarity between x_i and x_j. With the row-pivot view of W, the Euclidean distance between image x_i and x_j is computed

by $\|W_i - W_j\|_2$. Thus the distortion between W and prior similarity S^x can be computed by

$$\begin{aligned}\Omega(W, S^x) &= \frac{1}{2}\sum_{i,j=1}^{n} S_{ij}^x \left\| \frac{W_i}{\sqrt{d_i}} - \frac{W_j}{\sqrt{d_j}} \right\|_2^2 \\ &= \text{tr}\,(W^\top L^x W) = \text{tr}\, L^x (W W^\top) \end{aligned} \quad (1)$$

where $d_i = \sum_{j=1}^{n} S_{ij}^x$ is engaged for normalization purpose, L^x is the normalized graph Laplacian defined as

$$L^x = I - D^{-1/2} S^x D^{-1/2}, \quad (2)$$

where $D = \text{diag}(d_1, d_2, \ldots, d_n)$ is a diagonal matrix.

Similarly, we also compute the graph Laplacian L^t in concept space, and calculate the distortion between W and the textual similarity S^t:

$$\begin{aligned}\Omega(W, S^t) &= \frac{1}{2}\sum_{i,j=1}^{n} S_{ij}^t \left\| \frac{W_i^c}{\sqrt{d_i}} - \frac{W_j^c}{\sqrt{d_j}} \right\|_2^2 \\ &= \text{tr}\,(W L^t W^\top) = \text{tr}\, L^t (W^\top W) \end{aligned} \quad (3)$$

In the above, unlike the row-pivot view where the similarity induced by W in visual space is computed by $WW^\top$, the similarity in concept space is computed by $W^\top W$ under the column-pivot view.

Combining the above two-view formulations, we are ready to expose the objective function to address the first motivation, i.e., we should minimize the overall distortion between W and the two-view data:

$$\min_{W \in \mathbb{R}_+^{n \times m}} \lambda_x \text{tr}\, L^x (W W^\top) + \lambda_t \text{tr}\, L^t (W^\top W) \;\; : \;\; W_{ij} \geq 0 \quad (4)$$

where λ_x and λ_t controls the trade-off between visual information and concept information.

For the second motivation, we bound the difference between W and some initial relevance score T. Here we employ the the Frobenius norm:

$$\min_{W \in \mathbb{R}_+^{n \times m}} \|W - T\|_F^2 \;\; : \;\; W_{ij} \geq 0, W_{ij} = 0, \forall (i,j) \notin \mathcal{M} \quad (5)$$

where T is the initial weight matrix determined by users or some other ranking model, $\mathcal{M}$ denotes the indices of nonzeros in T, i.e., $M := \{(i,j) | T_{ij} \neq 0, i \in \mathbb{N}_n, m \in \mathbb{N}_m\}$.

Therefore, we obtain the following optimization by combining (4) and (5):

$$\begin{aligned} \min_W \; & \lambda_x \text{tr}\, L^x (W W^\top) + \lambda_t \text{tr}\, L^t (W^\top W) + \|W - T\|_F^2 \quad (6) \\ \text{s.t.} \quad & W \in \mathbb{R}_+^{n \times m}, W_{ij} \geq 0, \forall i \in \mathbb{N}_n, j \in \mathbb{N}_m, \\ & W_{ij} = 0, \forall (i,j) \notin \mathcal{M} \end{aligned}$$

So far we have the unified two-view framework. It is similar to two-view learning algorithms (see [6]) in the sense that the learning in the two views regularize each other such that the resultant solution is more robust. Comparing with previous work (for example, [21][12]) on tag ranking/refinement, it does not involve any probabilistic models. Probably it has more flexibility to fit the diverse data and avoids the difficulty in model selection.

3.4 Algorithm

There is no off-the-shelf optimization tools to solve (6) directly. Inspired by the ideas of sequential minimization [26], we propose to resolve (6) by an iterative projection algorithm, which is a variant of stochastic coordinate descent optimization [18].

In particular, in each optimization step, we randomly choose one row W_i to optimize and fix the rest $n-1$ rows. Consequently, the objective is simplified to be:

$$\begin{aligned} \min_{W_i} \quad & W_i (\lambda_t L^t + (\lambda_x L_{ii}^x + 1) I) W_i^\top \\ & + 2\lambda_x \sum_{j=1, j \neq i}^{m} L_{ij}^x W_j W_i^\top - 2 T_i W_i^\top \end{aligned} \quad (7)$$

where the constraints are omitted in the above formulation. This is a standard quadratic program [2] over vector W_i, which can be solved by the interior-point algorithm with typical polynomial time complexity of $O(m^3)$. Fortunately, we can derive the closed-form solution of (7) by dropping the constraints. Let J abbreviate the objective function (7), by taking its derivatives w.r.t. W_i, we have

$$\nabla J = W_i (\lambda_t L^t + (\lambda_x L_{ii}^x + 1) I) + \lambda_x \sum_{j=1, j \neq i}^{n} L_{ij}^x W_j^\top - T_i$$

Setting ∇J to 0 yields

$$W_i = (T_i - \lambda_x \sum_{j=1, j \neq i}^{n} L_{ij}^x W_j^\top)(\lambda_t L^t + (\lambda_x L_{ii}^x + 1) I)^{-1}, \quad (8)$$

which is the optimal solution of current W_i. Therefore we are making progress towards a local optimal objective value at each iteration step.

Despite the nice closed-form solution above, in practice, it remain challenging to directly compute the matrix inverse, which often has the time complexity of $O(m^3)$ for a dense matrix. This is because the size of tag vocabulary could be potentially very large in a real application; as a result, the computation of matrix inverse is prohibitive for large-scale applications. To overcome this obstacle, we use the Taylor approximation for the matrix inverse problem $(I + A)^{-1}$:

$$(I + A)^{-1} = I + \sum_{i=1}^{\infty} (-1)^i A^i.$$

As a result, we arrive at the approximate solution:

$$\begin{aligned} & (\lambda_t L^t + (\lambda_x L_{ii}^x + 1) I)^{-1} \approx \\ & \frac{1}{\lambda_x L_{ii}^x + 1} \left(I + \sum_{j=1}^{p} (-1)^j \left(\frac{\lambda_t}{\lambda_x L_{ii}^x + 1} L^t \right)^j \right) \end{aligned}$$

where p is the order of approximation. In practice, we can pre-compute the power of L^t and cache it for improving the time efficiency.

The solution (8) may violate the constraints over W. We project the solution into the feasible domain at each step:

$$W_{ij} = 0 \;\; \text{if } W_{ij} < 0 \text{ or } (i,j) \notin \mathcal{M}.$$

It implies we only need to consider the nonzero indices of W. Since each image is annotated by a very limited number of tags on average, W is essentially very sparse. Therefore the computation in (8) could be very efficient. Finally, we summarize the iterative projection solution in Algorithm 1.

Once the optimal tag weighting matrix W is obtained by the proposed algorithm, we rank the tags for an image x_i according to their relevance scores W_i's. For any two tags t_{ij} and t_{ik}, tag t_{ij} ranks on top of tag t_{ik} i.f.f $W_{ij} \geq W_{ik}$.

Algorithm 1 The Two-View Tag Weighing Algorithm.

Input: Social image corpus Z;
kernel function $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and $k_t : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_+$;
parameters λ_x, λ_t, p;
Output: Weighting matrix W.

```
Cluster the images into k groups;
for each group do
    Construct graph Laplacian L_t and L_x using k_t and k_x, respectively;
    Initialize W^0 = T
    repeat
        Randomly choose a row index i to update
        Compute U^j = (λ_t / (λ_x L^x_ii + 1) L^t)^j
        W_i = (T_i − λ_x Σ_{j=1,j≠i}^n L^x_ij W_j^⊤)(I + Σ_{j=1}^p (−1)^j U^j)/(λ_x L^x_ii + 1)
    until convergence criterion satisfied
end for
Tile the W in all groups to obtain W;
```

3.5 Similarity Measure for Building Graphs

The graph Laplacian L^x and L^t encode the local geometric information in visual space and concept space, respectively. It is our prior knowledge about the data distribution. With a proper similarity matrix S, the graph Laplacian can be computed immediately from (2). Therefore we just focus on S^x and S^t.

3.5.1 Similarity in Visual Space

Let $[S^x]_{ij} := k_x(x_i, x_j)$ be the similarity matrix of images computed from some function $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Depending on the features employed to represent a feature, one can adopt different kinds of similarity functions to define k_x. How to extract features for representing an image remains a very challenging problem itself. We will discuss our approach in the experimental section. Here we discuss how to compute the similarity between two images for two major types of features: global and local features.

For global features, a typical approach for similarity measure is based on a Gaussian kernel::

$$k_x^g(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2}). \tag{9}$$

In this case the visual space $\mathcal{X}$ can be deemed as a subset of $\mathbb{R}^d$, where d is the number of extracted features. We present a set of global features in Section 4.

For local features (such as SIFT [14]), each image is represented by a bag of descriptors. To measure similarity between two images given the two bags of descriptors, we employ a simple yet effective matching kernel [19] as follows:

$$k_x^l(x_i, x_j) = \frac{1}{2}(\hat{k}(x_i, x_j) + \hat{k}(x_j, x_i)) \tag{10}$$

where $\hat{k}(x_i, x_j)$ is defined as

$$\hat{k}(x_i, x_j) = \frac{1}{|x_i|}\sum_{s=1}^{|x_i|} \max_t \tilde{k}(d_{i,s}, d_{j,t}) \tag{11}$$

where $\tilde{k}(d_{i,s}, d_{j,t})$ measures similarity between two descriptors. Finally, we combine both the global and the local similarity functions for computing similarity in visual space.

3.5.2 Similarity in Concept Space

Given a collection of social images, we can mine the relationship among tags by carefully studying the related statistics. For example, if two tags appear together frequently, the distance between them should be reasonably small. Here we adopt the Google distance [3]:

$$d(t_i, t_j) = \frac{\max(\log f(t_i), \log f(t_j)) - \log f(t_i, t_j)}{\log n - \min(\log f(t_i), \log f(t_j))},$$

where $f(t_i)$ is the number of images containing tag t_i, $f(t_i, t_j)$ is the number of images containing both t_i and t_j.

Our target is the similarity matrix S^t. With the Google distance function, we compute the similarity among tags by

$$k_t(t_i, t_j) = \exp(-d(t_i, t_j)). \tag{12}$$

For more other methods to explore the correlation among tags, please refer to [9] for more examples.

3.6 Speedup by Clustering

The updating of W is benefited from the sparseness of the annotated tags. However, in real application, the corpus size n and the vocabulary size m could be very large, which makes the algorithm slow. We stress the fact that social images form meaningful groups. Images from different groups share little similarity. Thus we can employ some clustering algorithm to separate the images into g visual groups. The inter-group similarity is simply set to zero.

In order to upper bound the size of each cluster effectively, we employ a bisecting clustering algorithm [27]. At each iteration, the largest group is chosen to split. The clustering objective function is to maximize the overall inner-cluster similarity [27], i.e.,

$$I_2 = \sum_{k=1}^{g} \sum_{x_i \in S_k} cos(x_i, C_k),$$

where g is the target number of groups, S_k is the k-th group, and C_k is the centroid of S_k. By this strategy, the size of the Laplacian L^x reduces from n to a small number, and thus the overall computational cost can be reduced.

4. EXPERIMENTS

We conduct an extensive set of experiments to verify our tag ranking algorithm, and apply our technique to two important applications: text-based social image retrieval and automatic tag recommendation.

4.1 Experimental Testbed and Setup

We crawled a data set consisting of about 1,000,000 social images from Flickr [2]. To evaluate the proposed algorithm, we form the evaluation testbed by choosing a number of query images that are related to a wide range of tags, including animals, plants, humans, landmarks, natural sceneries, and some daily objects.

For feature representation, we extract four kinds of effective global features by [28]. These features include: (1) 81-dimensional grid color moment features, (2) 59-dimensional Local Binary Pattern (LBP) texture features [16], (3) 120-dimensional Gabor wavelets texture features [10], and (4) 37-dimensional edge direction histogram features. In total, a 297-dimensional vector is used to represent an image in the

[2]http://www.flickr.com/

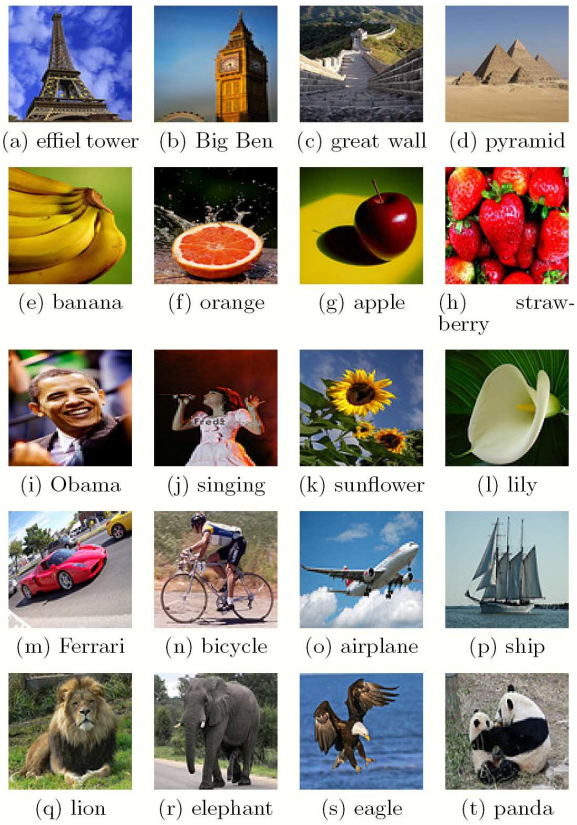

Figure 3: Illustration of sample figures related to the queries in our experiments.

data sets. The similarity of global features in visual space in computed by the Gaussian kernel in Eq (9). For local features, we employ SIFT feature descriptors [14] to extract local features and employ the kernel function in Eq. (10) to define the similarity of local features in visual space. A simple linear combination is adopted to combine the similarity of both global and local features.

For the preprocess on tags, we first filter out the extremely unpopular tags. Then we adopt the Gaussian with Google distance to calculus the similarity in concept space (12).

For experimental setting, there are several hyper-parameters used in our two-view tag ranking algorithm. For most of our results, these parameters are set below:

- σ: the band-width parameter of Gaussian kernel. We set it to be the average square Euclidean distance among all the images;
- λ_x and λ_t: the trade-off parameters in (6). We set $\lambda_x = 0.5$ and $\lambda_t = 1.0$ empirically;
- g: the number of groups in the clustering process. We set it to 120. We employ the *CLUTO* toolkit [3] to help to do clustering.
- p: the approximation order when computing the matrix inverse. We use $p = 5$.
- T: the initial tag weight matrix. For an image x_i, we set its tags the uniform weight $1/|t_i|$, where $|t_i|$ is the number of tags annotated to x_i.

[3]http://glaros.dtc.umn.edu/gkhome/views/cluto/

Table 1: The statistics of the queries in the data set.

TestQuery	#RelDoc	TestQuery	#RelDoc
eiffel tower	919	Barack Obama	88
great wall	57	Big Ben	200
red car	477	pyramid	229
airplane	771	Ferrari	401
lily	452	bicycle	1475
banana	124	sunflower	798
fruit orange	266	strawberry	715
singing	945	panda	1512
lion	1308	sheep	487
elephant	739	eagle	664

4.2 Performance Evaluation on Tag-based Social Image Retrieval

We focus on the *query-by-text* setting. We assume ground truth rankings have two grades, *relevant* and *irrelevant*. In order to make the results reliable, we pick up 20 queries and employ 6 staffs to label the images according to whether they feel them relevant to the queries or not. We choose the queries that are popular and diverse. The complete list of queries are shown in Table 1.

4.2.1 Ranking Schemes and Evaluation Metric

For an input query q, we only consider the social images containing q as candidate relevant images. For an image x_i, let π_i denote the position of q in the ranked tag list of x_i. We define the relevance score in the way of [12]:

$$r(x_i) = -\pi_i + \frac{1}{n_i} \tag{13}$$

Note two key properties of this scheme: 1) if $\pi_i < \pi_j$, we always have $r(x_i) > r(x_j)$, which means we assign higher relevance score to the image containing the query tag at more advanced positions in its ranked tag list; 2) if $\pi_i = \pi_j$, the image having fewer tags is assigned larger relevance score. The motivation is that the more tags an image possesses, the more noisy of its visual content.

To our knowledge, [12] is the first and only paper aiming to attack the tag ranking problem. However, we are aware that some existing image annotation or annotation refinement works can be easily adapted to generate a permutation over the tags. For example, CMRM[8], WNM[9], RWRM[20] and CIAR[21]. One can always rank the tags according to their relevance score to the image. Since it has been shown that CIAR outperforms the other annotation methods, we evaluate it specifically for comparison purpose. To summarize, we evaluate the following tag-ranking schemes:

- **Baseline**: the original order of the tags is maintained, i.e., the tag position is determined by some web users;
- **CIAR**: adaption of the Content-Based image Annotation Refinement algorithm in [21]. They estimate the condition probability $p(t_j|x_i)$ of generating tag t_j by image x_i by Gaussian kernel density. We use this score to rank tags;
- **GM-RW**: the method proposed by [12]. The tag ranking score is first computed by a Generative Model

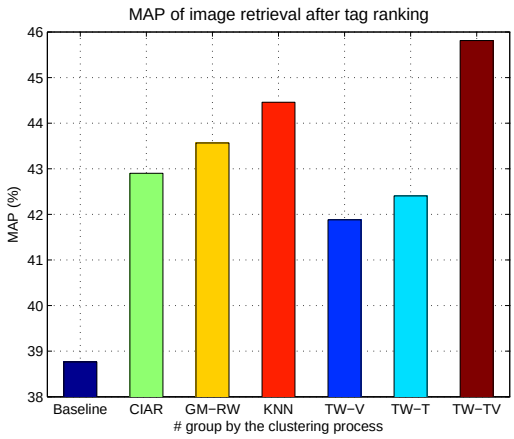

Figure 4: Mean average precision of different tag-based image retrieval methods after tag ranking.

$p(x_i|t_j)$ and then perform Random Walk over a similarity graph on tags to refine the score;

- **KNN**: the method proposed by [11]. The tag ranking score is the total number of votes received from its visual nearest neighbors. It is shown in [11] that this voting method could be better than model-based methods in [21];
- **TW-TV**: the proposed two-view Tag Weighing method that combines the local information both in Tag space and in Visual space. The objection function is (6). All the hyper-parameters are described in Section 4.1. The solution is obtained from Algorithm 1;
- **TW-V**: the proposed two-view Tag Weighing method in Visual space, that is, $\lambda_t = 0$ in objective (6);
- **TW-T**: the proposed model-free Tag Weighting method in Tag space, that is, $\lambda_x = 0$ in objective (6); .

To evaluate the performance, we use the standard *mean average precision* (MAP) measure [1]: Let π^* be the ground-truth ranking and π be the ranking result by some relevance score r, the average precision score is defined as

$$MAP(\pi^*, \pi) = \frac{1}{rel} \sum_{j:\pi_j^*=1} Prec@j,$$

where $rel = |\{i : \pi_i^* = 1\}|$ is the number of relevant documents, and $Prec@j$ is the percentage of relevant documents in the top j documents n predicted ranking r. MAP is the mean of the average precision scores of a group of queries.

4.2.2 Results of Image Retrieval Accuracy

We plot the MAP measure in Figure 4.2.2.

First we observe that all the tag ranking methods outperforms the baseline method significantly. This verifies the necessity of tag ranking. It also coincides with the statements of [12] that the relevant tags may not be ranked at the top positions. Thus the ranking score computed by (13) cannot reflect the relevance of the image to the query tag effectively.

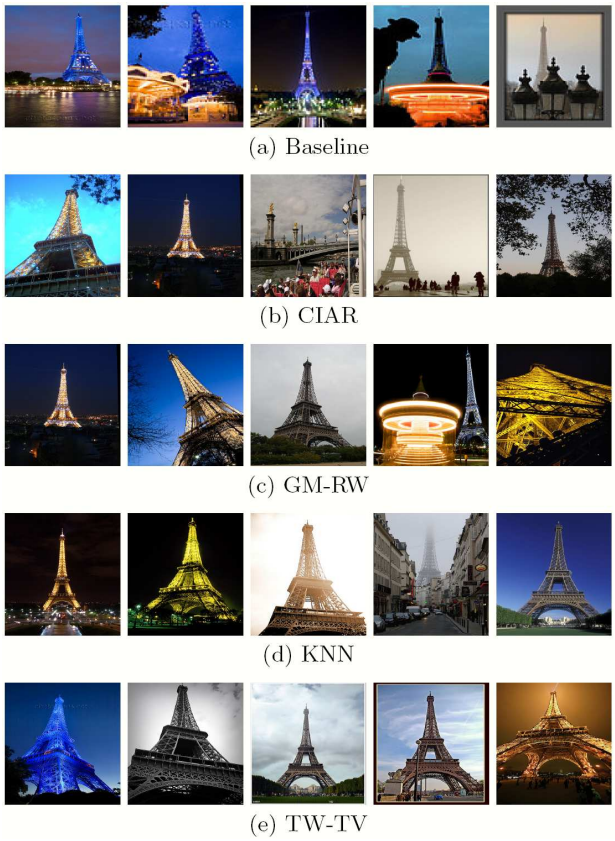

Figure 5: The top 5 tag-base retrieval results on the query eiffel tower after different tag ranking algorithms. One can see that the result of TW-TV is visually more pure.

Second, we observe that our two-view tag ranking scheme works well. It produces very competitive results with CIAR, GM-RW, and KNN (actually TW-TV is the best on our data set). Both CIAR and GM-RW involves generative models and performs a random walk over the similarity graph S^t over tags. We observe that GM-RW is better than CIAR. GM-RW makes use of both exempler-based visual information and statistical information of tags when constructing S^t, while CIAR only uses visual information. This probably explains the advantage of GM-RW. However, a simple k nearest neighbor voting method outperforms model-based CIAR and GM-RW. This implies that when large amount of data is accessible, we should pay more effort to data-driven approaches. The best algorithm TW-TV in Figure 4.2.2 is a model-free learning scheme. Without assume any parametric model between visual and tag spaces, it has more flexibility to fit the diverse data, which is our motivation of the two-view learning algorithm. We also stress that both TW-TV and GM-RW explores the same source of information. However, our method is conceptually simple and purely data-driven. It has fewer hyper-parameters and can be easily to large-scale application. We show this computation merit of TW-TV in next section.

At last, the combination of two view is better than using single view. The membership between images and tags con-

nects the visual view to concept view. The learning in each view regularizes each other such that the resultant solution is more robust. The essential complementary property of visual and tag space makes TW-TV more effective than other data-driven method KNN.

Finally, Figure 5 shows some examples to examine the qualitative performance of the retrieval results achieved by different methods.

4.3 Evaluation of Computation Efficiency

In this section, we examine the time complexity of the TW-TV and GM-RW tag ranking methods over a randomly chosen subset consisting of 10,000 images. For both methods, we pre-compute the similarity matrix in visual space. The tag similarity based on Google distance is also pre-computed and is excluded from the running time reported here. For the proposed TW-TV method, we cluster the images into groups such that the resultant maximal group has no more than 2,000 images. Our goal is to test the time complexity with the increment of image number n.

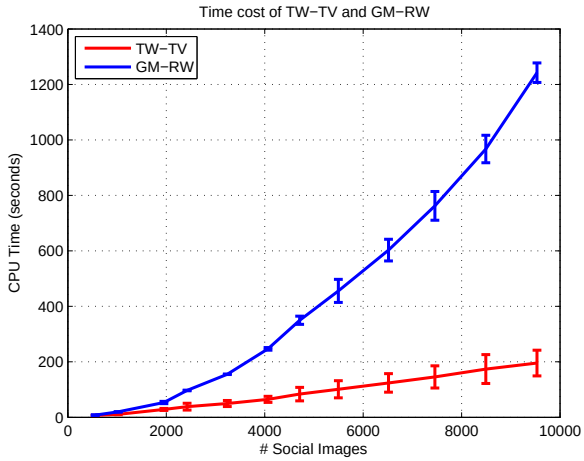

Figure 6: Time cost of TW-TV and GM-RW tag ranking algorithms with the number of social images. The results are averaged over 10 times run.

We plot the CPU time in Figure 6. It is clear to observe that GM-RW is significantly slower than TW-TV for all n values. When $n = 10,000$, GM-RW took about 20 minutes, while TW-TV took only about 3 minutes. More importantly, the curve of GM-RW shows that it empirically has the polynomial complexity on n, while TW-TV exhibits asymptotic linear time cost. To measure this more quantitatively, we fit the curves in Figure 6 by least square fitting. The estimated empirical complexity of TW-TV and GM-RW are $O(n^{0.48})$ and $O(n^{0.63})$, respectively. This efficiency gain is crucial for real systems dealing with large-scale data set.

The empirical linear time complexity of TW-TV can be interpreted by the clustering process in Section 3.6. By dividing the data set into small groups, n drops to the size of some single cluster, which is a much smaller than the original one. Therefore, the overall time cost is linear on the number of clusters, even though the total number of images could be potentially huge. Despite the encouraging results, we note that we do not count the clustering time complexity, which depends on the applied clustering algorithms.

4.4 Evaluation of Hyper-parameters

The proposed TW-TV tag ranking algorithm has a few free parameters. The most important ones may be λ_x and λ_t which control the trade-off between the importance of visual information and tag statistics. Note that we deem all the annotated tags have identical prior weight for any image. Thus T would not affect the tag weighing result. If we scale λ_x and λ_t by the same value, the final weight W remains the same. Only the ratio $\frac{\lambda_x}{\lambda_t}$ affects the performance. So we restrict λ to the range $[0, 1]$. The MAP results are plotted in Figure 7.

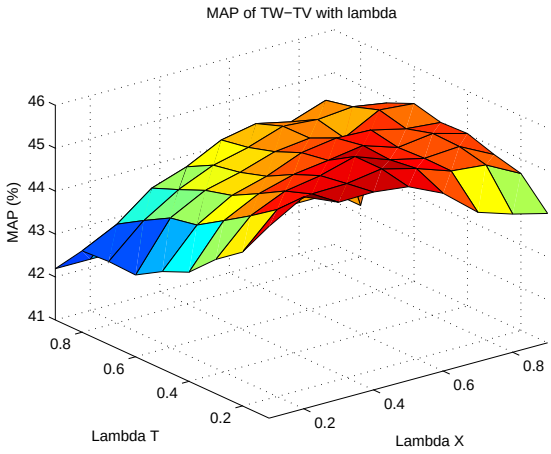

Figure 7: MAP of tag-based image retrieval after TW-TV with different hyper-parameter λ_x and λ_t. T is set to be uniform value for each image. Only the ratio λ_x/λ_t affects the performance.

First of all, we see that the retrieval performance is not sensitive to λ in the range $[0.4, 0.8]$. The fluctuation of MAP is bounded by 2 percentage. Moreover, the MAP measure varies smoothly with λ. This shows the robustness of our method, which is probably induced by the complementary information in both spaces. At the center of the surface, TW-TV produces the best MAP. This is somewhat surprising because empirically visual information is more important. For example, Liu et. al. [12] set the ratio of visual and tag importance as 4:1. We conjecture that maybe this is resulted by the image representations. We extract different visual features with [12]. It also suggests that there are potentials to improve the performance of TW-TV. When λ_x/λ_t is skew, that is, the most left and the most right square of the surface, the performance is poor. It means that single view cannot result in satisfactory solution.

The success of TW-TV relies on the local geometric information of image and tag similarity graphs. We take the following 3 steps to build the Laplacian in visual space: for each image, 1) using global and local features to locate the k nearest neighbors; 2) in the similarity graph, set the value at these k-NN indices to 1, otherwise 0; 3) compute the normalized Laplacian by (2). The similarity in tag space is

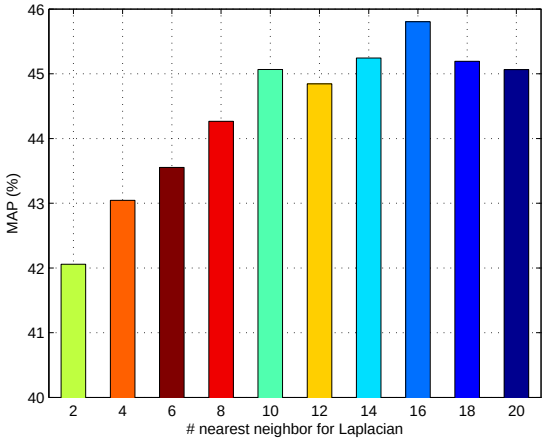

Figure 8: MAP of TW-TV with different number of nearest neighbors when building graph Laplacian in visual space.

constructed using Google distance. We present the influence of the value k in Figure 8.

We observe that the MAP increases with k when k is small. When k is larger than 10, the increment in MAP is marginal. The most possible reason is that the local geometry is already preserved for reasonably large k. Moreover, we see when $k = 20$, the performance actually starts to drop. In this case, the images that do not share much similarity is considered as neighbors. Therefore, such images are essentially noise and misleads the learning. The MAP measure is not too sensitive to k.

The clustering process plays an important role to make the whole weighting method practical. We evaluate the MAP with different group number g in Figure 9.

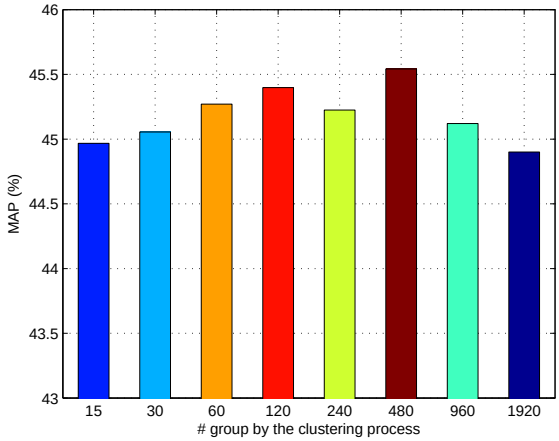

Figure 9: MAP of TW-TV with different number of groups in the clustering process.

From equation (8), we see the quadratic complexity over n. Therefore, for a fixed image corpus, the more groups we have, the faster of TW-TV. On the other hand, inter-group similarity is ignored in our algorithm. Therefore we should be aware of the granularity of the groups. If g is too small, we could loss the local similarity information. The above figure shows the trend of MAP with the number of groups g. It is shown that MAP is not sensitive to g. For g ranges from 15 to 2000, the MAP value fluctuates within 1 percentage. Thus we conclude the clustering process is reasonable and effective for TW-TV. When g is greater than 480, MAP starts to drop. This verifies our conjecture about the granularity of the clustering.

4.5 Application to Auto Tag Recommendation

After we have the tag ranking results, we can use the images with properly ordered tags to annotate a given image automatically. We need not require users to provide any initial candidate tags. We implement this by a simple nearest neighbor voting method, first proposed in [12]. For an input image, we first locate its k nearest neighbors in the corpus. Then we extract the top h tags of each of these neighbors. So have a collection of size $k \times h$. Then we remove the redundant ones in this collection and recommend the resulting tag set to the image. For each tag in this collection, we can indicate its weight by this redundancy number. Accordingly we obtain a rank of the annotated tags.

Figure 10: An illustration of image annotation results with the TW-TV tag ranking scheme.

Figure 10 shows some examples by applying the proposed TW-TV tag ranking method. It is clear to observe that most of the top ranked tags are quite relevant to the semantic concepts of the query images.

5. LIMITATIONS AND DISCUSSIONS

Despite the encouraging results achieved above, there are still several limitations for our study in this paper. First of all, we only exploit textual tags and visual contents of social images for ranking the tags of social images. However, for real-world social images, there often exists other user-generated contents, such as descriptions, user info, rating, and geo-tags, etc. In future work, we plan to extend our method by adding the extra information, which could further improve the performance of the tag ranking scheme.

Moreover, in the proposed scheme, we apply a simple clustering processing step to speed up the solution in order to handle large data, which may lose some information in some situation. In future work, we plan to investigate more sophisticated techniques to improve the efficiency of our method.

6. CONCLUSIONS

We propose a novel two-view learning approach to address the tag ranking problem in this paper. We do not assume any parametric model on the relevance between images and tags. It is a pure data-driven algorithm, which distinguishes it from previous generative model based works on this topic. Therefore it could probably have better flexibility to fit the diverse data. The key idea is inspired by two-view learning algorithms. When the representation of data has two complementary views, we can make use of the information of each to regularize each other such that the resultant solution is more robust. We also devise efficient algorithm and practical strategies to speedup the learning. Empirical results on a real data set exhibits both the efficacy and efficiency of our methods. Future work will extend our method to exploit other contents of social images for tag ranking.

Acknowledgement

This work was supported by Singapore MOE tier-1 grant (RG67/07).

7. REFERENCES

[1] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] R. Cilibrasi and P. M. B. Vitányi. The google similarity distance. *IEEE Trans. Knowl. Data Eng.*, 19(3):370–383, 2007.

[4] R. Datta, W. Ge, J. Li, and J. Z. Wang. Toward bridging the annotation-retrieval gap in image search by a generative modeling approach. In *ACM Multimedia*, pages 977–986, 2006.

[5] L. Y. X.-S. H. H. J. Z. Dong Liu, Meng Wang. Tag quality improvement for social images. In *Multimedia and Expo*, pages 350–353, 2009.

[6] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmák. Two view learning: Svm-2k, theory and practice. In *NIPS*, 2005.

[7] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. *CoRR*, abs/cs/0508082, 2005.

[8] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR*, pages 119–126, 2003.

[9] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & wordnet. In *ACM Multimedia*, pages 706–715, 2005.

[10] M. Lades, J. C. Vorbrüggen, J. M. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers*, 42(3):300–311, 1993.

[11] X. Li, C. G. M. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Multimedia Information Retrieval*, pages 180–187, 2008.

[12] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *WWW*, pages 351–360, 2009.

[13] X. Liu, R. Ji, H. Yao, P. Xu, X. Sun, and T. Liu. Cross-media manifold learning for image retrieval & annotation. In *Multimedia Information Retrieval*, pages 141–148, 2008.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[15] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, 2006.

[16] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[17] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW*, pages 327–336, 2008.

[18] J. C. Spall. *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, Inc., New York, NY, USA, 2003.

[19] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *ICCV*, page 257–264, Washington, DC, USA, 2003.

[20] C. Wang, F. Jing, L. Zhang, and H. Zhang. Image annotation refinement using random walk with restarts. In *ACM Multimedia*, pages 647–650, 2006.

[21] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Content-based image annotation refinement. In *CVPR*, 2007.

[22] C. Wang, L. Zhang, and H.-J. Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In *SIGIR*, pages 355–362, 2008.

[23] K. Q. Weinberger, M. Slaney, and R. van Zwol. Resolving tag ambiguity. In *ACM Multimedia*, pages 111–120, 2008.

[24] L. Wu, S. C. Hoi, J. Zhu, R. Jin, and N. Yu. Distance metric learning from uncertain side information with application to automated photo tagging. In *Proceedings of ACM International Conference on Multimedia (MM2009)*, Beijing, China, Oct. 19–24 2009.

[25] L. Wu, L. Yang, N. Yu, and X.-S. Hua. Learning to tag. In *WWW*, pages 361–370, 2009.

[26] T. Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682–691, 2003.

[27] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, 2004.

[28] J. Zhu, S. C. H. Hoi, M. R. Lyu, and S. Yan. Near-duplicate keyframe retrieval by nonrigid image matching. In *ACM Multimedia*, pages 41–50, 2008.