

Image Annotation by k NN-Sparse Graph-based Label Propagation over Noisily-Tagged Web Images

JINHUI TANG, RICHANG HONG, SHUICHENG YAN, TAT-SENG CHUA

National University of Singapore

GUO-JUN QI

University of Illinois at Urbana-Champaign

and

RAMESH JAIN

University of California, Irvine

In this paper, we exploit the problem of annotating a large-scale image corpus by label propagation over noisily-tagged web images. To annotate the images more accurately, we propose a novel k NN-sparse graph-based semi-supervised learning approach for harnessing the labeled and unlabeled data simultaneously. The sparse graph constructed by datum-wise one-vs- k NN sparse reconstructions of all samples can remove most of the semantically-unrelated links among the data, and thus it is more robust and discriminative than the conventional graphs. Meanwhile, we apply the approximate k nearest neighbors to accelerate the sparse graph construction without losing its effectiveness. More importantly, we propose an effective training label refinement strategy within this graph-based learning framework to handle the noise in the training labels, by bringing in a dual regularization for both the quantity and sparsity of the noise. We conduct extensive experiments on a real-world image database consisting of 55,615 Flickr images and noisily-tagged training labels. The results demonstrate both the effectiveness and efficiency of the proposed approach and its capability to deal with the noise in the training labels.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing-indexing methods; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

General Terms: Algorithms, Theory, Experimentation

Additional Key Words and Phrases: Sparse Graph, k NN, Semi-supervised Learning, Label Propagation, Web Image, Noisy Tags

Author's address: J. Tang, R. Hong, and T.-S. Chua, School of Computing, 13 Computing Drive, 117417, Singapore. S. Yan, Department of Electrical and Computer Engineering, 4 Engineering Drive 3, 117576, Singapore. G.-J. Qi, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL 61801, USA. R. Jain, Bren School of Information and Computer Sciences, University of California, Irvine, CA 92697, USA.

Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.

© 2010 ACM 1529-3785/2010/0700-0111 \$5.00

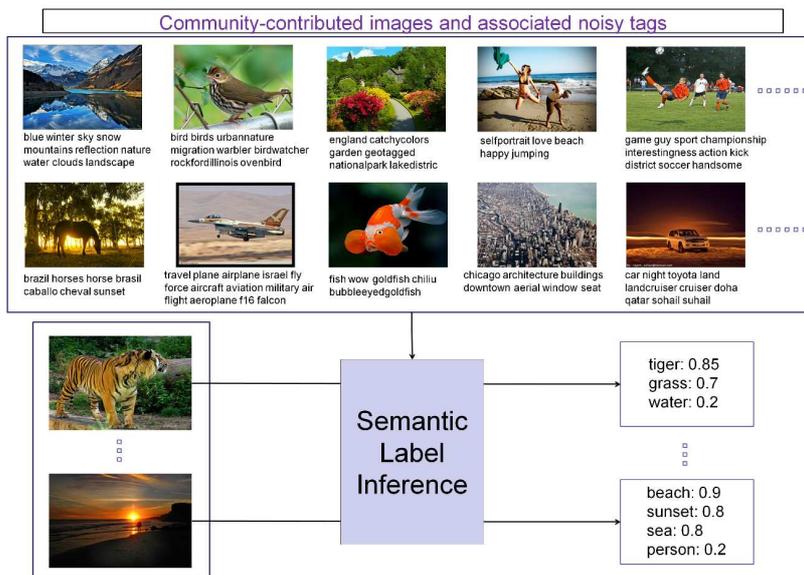


Fig. 1. An overview of the process of inferring images' labels from community-contributed images and noisy tags.

1. INTRODUCTION

Recent years have witnessed the proliferation of social media and the success of many photo-sharing websites, such as Flickr¹ and Picasa². These websites allow users to upload personal images and assign tags to describe the image contents. With the rich tags as metadata, users can more conveniently organize and access these shared images. Out of these, a question naturally arises for research on image annotation (i.e., inferring images' labels on different semantic concepts) — can we infer the images' semantic labels effectively from these user-shared images and their associated tags?

Utilizing machine learning techniques to improve image annotation performance has attracted much attention in the multimedia research community. However, the effectiveness of these machine learning techniques heavily relies on the availability of a sufficiently large set of balanced labeled samples, which typically come from users in an interactively manual process. This manual labeling process is very time-consuming and labor-intensive. In order to reduce this manual effort, many semi-supervised learning or active learning approaches have been proposed [He et al. 2004][Goh et al. 2004]. Nevertheless, there is still a need to manually annotate a large set of images to train the learning models. On the other hand, the image sharing sites offer us a great opportunity to “freely” acquire a large number of images with annotated tags. The tags of the images are collectively annotated by a large group of heterogeneous users. It is generally believed that most of the tags

¹<http://www.flickr.com/>

²<http://picasa.google.com/>

are correct, although there are many incorrect and missing tags. Thus if we can infer the images' semantic labels effectively from these user-shared images by using their associated noisy³ tags as training labels, considerable manual efforts can be eliminated for labeling the training data. Fig. 1 presents an overview of the process of inferring images' labels from the community-contributed images and associated noisy tags.

Many traditional methods, such as the support vector machine and k nearest neighbors (k NN) method, can be applied to infer the images' labels from the user-shared images and associated tags. To annotate the images more accurately, we propose a novel k NN-sparse graph-based semi-supervised learning method in this paper. By utilizing the labeled and unlabeled data simultaneously, semi-supervised learning has been demonstrated to be more effective than purely supervised learning when the training data is limited [Chapelle et al. 2006][Zhu 2005]; and graph-based semi-supervised learning methods have been widely used in image annotation and retrieval [He et al. 2004][Wang et al. 2006]. However, the traditional graph-based methods share a common disadvantage, namely, they all have certain parameters which require manual tuning. The parameters may have great impact on the structure of the constructed graph. Thus the performances of these methods are sensitive to these parameters and the algorithmic robustness is challenged. Meanwhile, the traditional graphs are constructed only based on visual distance, and there may exist many links between the samples with unrelated concepts. This may cause the label information to be propagated incorrectly.

Actually a graph is a gathering of pairwise relations, while the relation among visual images is essentially an estimation based on human cognition system. It has been found in neural science that the human vision system seeks a sparse representation for the incoming image using a few words in a feature vocabulary [Rao et al. 2002]. This motivates us to construct the so-called sparse graph through the sparse reconstructions of the samples. However, the one-vs-all reconstruction is very time-consuming and needs too much memory for large-scale dataset. Thus compared to our previous work [Tang et al. 2009], we sparsely reconstruct each sample from its k nearest neighbors in feature space instead of using all the other samples to improve the efficiency while maintaining its effectiveness. We call it one-vs- k NN sparse reconstruction. The approximate k NN search [Mount and Arya 1997] is also applied to accelerate the process. The semi-supervised label inference for semantic concepts is then conducted on this sparse graph.

The k NN-sparse graph-based semi-supervised learning method has the following advantages: 1) it can remove most of the semantically-unrelated links to avoid the propagation of incorrect information, since each sample only has links to a small number of most probably semantically-related samples; 2) it is robust to noisy elements in the visual features; 3) it is naturally effective for discrimination since the sparse graph characterizing the local structure can convey important information for classification [Belkin and Niyogi 2003]; and 4) it is practical for large-scale applications since the sparse representation can reduce the storage requirement while the approximate k NN-sparse graph construction is much more efficient than normal sparse graph construction.

³In this paper, we use "noisy" to denote both "incorrect" and "incomplete" ("missing") for tags.

More importantly, in this graph-based learning framework, we propose an effective training label refinement strategy to handle the noise in the training labels, by bringing in a dual regularization for both the quantity and sparsity of the noise. This is a key-point for our scenario, since our training labels are extracted from the community-contributed noisy tags.

We conduct extensive experiments on a real-world dataset [Chua et al. 2009] consisting of 55,615 community-contributed images and their associated tags crawled from Flickr to demonstrate the advantages of the proposed k NN-sparse graph-based semi-supervised learning approach and the training label refinement strategy for noisy tag handling.

The main contributions of this work are as follows:

- (1) We propose a novel k NN-sparse graph-based semi-supervised learning approach, which is more consistent with human perception. The sparse graph can remove most of the semantically-unrelated links among different samples to improve the effectiveness. We also apply the approximate k NN search to accelerate the sparse graph construction and reduce the storage requirement for large-scale applications without losing the effectiveness.
- (2) An effective training label refinement strategy is proposed within this graph-based learning framework to handle the noise in the tags, by bringing in the dual regularization for both the quantity and sparsity of the noise.

2. RELATED WORK

Several approaches have been proposed for annotating images by mining the web images with surrounding descriptions. A series of research were also done to leverage information in world-wide web to annotate general images [Wang et al. 2008][Wang et al. 2006][Li et al. 2006]. Given a query image, they first searched for similar images from the web, and then mined representative and common descriptions from the surrounding descriptions of these similar images as the annotation for the query image. By fully leveraging on the redundancy of information on the Web, [Torralba et al. 2008] collected about 80 million tiny images, each of which is labeled with one of the 75,062 abstract nouns from WordNet. They claimed that with sufficient number of samples, the simple nearest neighbor classifier can achieve reasonable performance for several object/scene detection tasks such as the human and face detection, when compared with the more sophisticated state-of-the-art techniques. However, the assignment of only one noun to each image and the use of small sized image of 32-by-32 pixels are inadequate to reflect the complex content of real-world images. The above efforts annotated the query image by collecting the descriptions of its similar images in the web. Their robustness to the noisy descriptions and the semantically unrelated neighbors are challenged.

Several approaches tried to model the visual patterns of certain concepts by mining the images gathered from the web. After web images are gathered using an object name, [Fergus et al. 2005] modeled the visual object as a constellation of parts using a probabilistic representation called TSI-pLSA. In [Sun et al. 2008], web images were mined to obtain multiple visual patterns automatically that are then used to model a semantic concept. Both these approaches gathered a sepa-

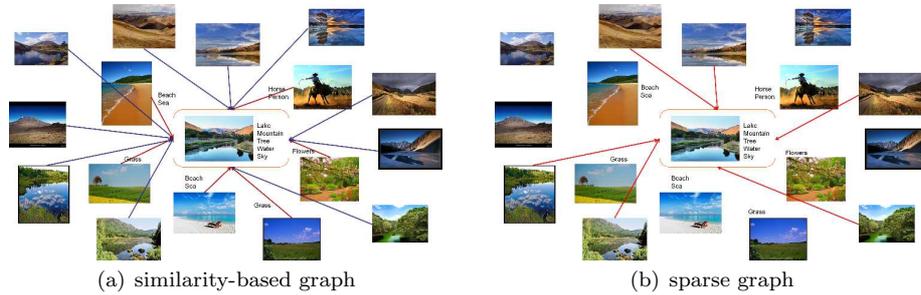


Fig. 2. The exemplary comparison of conventional similarity-based graph and sparse graph.

rate training set for each concept, and need to train models for different concepts separately. Thus their applications are limited to a small number of concepts.

Recently, graph-based semi-supervised learning has attracted much attention in both machine learning and multimedia retrieval communities. Actually graph-based semi-supervised learning is a label propagation process [Tang et al. 2007]. The most typical ones include the Gaussian random fields and harmonic functions method [Zhu et al. 2003], as well as the local and global consistency method [Zhou et al. 2003]. However, they both have the disadvantage of the requirement to tune certain parameters. Another popular method is the linear neighborhood propagation [Wang and Zhang 2008], in which the sample reconstruction method is used to construct a graph. It has been shown that in most cases, linear neighborhood propagation is more effective and robust than the traditional semi-supervised methods on similarity graphs [Wang and Zhang 2008][Tang et al. 2008]. However, it still cannot handle the links among semantically-unrelated samples. A more detailed survey on semi-supervised learning can be found in [Zhu 2005].

In this paper, a novel k NN-sparse graph-based semi-supervised learning method with regularization on training labels is proposed to annotate images by label propagation over the noisily-tagged web images. Here the graph is constructed sparsely to handle the semantically-unrelated links. It is constructed by reconstructing each sample from its k nearest neighbors to improve the efficiency, while the approximate method is applied to accelerate the k NN search. And the regularization is proposed to handle the noise in the training labels.

3. SEMI-SUPERVISED INFERENCE ON KNN-SPARSE GRAPH

Most traditional graph-based semi-supervised learning algorithms construct the graphs only according to the visual distance, thus are very sensitive to the noise in visual features. One dimension of noisy feature may affect the graph structure significantly. Moreover, constructing the graph only based on the visual distance will bring in semantically-unrelated links between samples due to the *semantic gap*. An alternative way to construct a graph is to reconstruct each image by the other images as in locally linear embedding [Roweis and Saul 2000] and linear neighborhood propagation [Wang and Zhang 2008]. However, they still cannot handle the semantically-unrelated links.

It has been found in neural science that the human vision system seeks a sparse

representation for the incoming image using a few words in a feature vocabulary [Rao et al. 2002]. [Wright et al. 2009] demonstrated that the ℓ_1 -norm based linear reconstruction error minimization can naturally lead to a sparse representation for the images. The sparse reconstruction is robust to the noise in features, and shows to enforce the images selected to reconstruct the test image are semantically-related to the test image. This motivates us to construct the graph by datum-wise sparse reconstructions of samples via ℓ_1 -norm minimization. The graph constructed by datum-wise sparse reconstruction of samples can remove considerable semantically-unrelated links between those semantically unrelated samples to avoid incorrect information propagation. However, the one-vs-all sparse reconstruction is computationally very complex. Thus compared to our previous work [Tang et al. 2009], we propose the so-called one-vs- k NN sparse graph construction by reconstructing each sample from its k nearest neighbors instead of reconstruction from all the other samples. Meanwhile, the one-vs-all sparse reconstruction needs too much memory to store the feature vectors of all samples for the optimization. Thus it is not feasible for the large-scale applications. While the one-vs- k NN sparse reconstruction can tackle this problem effectively as it only needs to store $k+1$ feature vectors for each reconstruction.

In addition, considering the fact that the training labels are noisy as they are extracted from the user-contributed tags, we propose a training label refinement strategy to restrain the effects of the noise in the training labels, by introducing a dual regularization for both the quantity and sparsity of the noise into the optimization of label inference.

Fig. 2 shows an exemplary comparison of sparse graph and conventional similarity-based graph. In the similarity-based graph, there exists link for each sample pair, and the weight is in inverse proportional to the distance measured in visual space. Thus the information may be propagated between semantically unrelated samples. In the constructed sparse graph, only a small number of most probably semantically-related samples are selected to have links to the reference sample. Thus the sparse graph can remove most of those semantically-unrelated links between images to avoid propagation of incorrect information.

3.1 k NN-Sparse Graph Construction

The pursue of the sparsest solution for sample reconstruction over an overcomplete dictionary is an NP-hard problem in general. However, recent results [Donoho 2006] show that if the solution is sparse enough, the sparse representation can be recovered by convex ℓ_1 -norm minimization. Suppose we have an under-determined system of linear equations: $\mathbf{x} = \mathbf{D}\mathbf{w}$, where $\mathbf{x} \in \mathbb{R}^d$ is the feature vector of the image to be reconstructed, $\mathbf{w} \in \mathbb{R}^n$ is the vector of the unknown reconstruction coefficients, and $\mathbf{D} \in \mathbb{R}^{d \times n}$ ($d < n$) is a matrix formed by the feature vectors of the other images in the dataset. The sparse solution for \mathbf{w} can be obtained by solving the following convex optimization problem [Donoho 2006]:

$$\min_{\mathbf{w}} \|\mathbf{w}\|_1, \quad s.t. \quad \mathbf{x} = \mathbf{D}\mathbf{w}. \quad (1)$$

In practice, there may exist noise on certain elements of \mathbf{x} , and a natural way to recover these elements and provide a robust estimation of \mathbf{w} is to formulate $\mathbf{x} = \mathbf{D}\mathbf{w} + \xi$, where $\xi \in \mathbb{R}^d$ is the noise term. We can then solve the following

ℓ_1 -norm minimization problem with respect to both reconstruction coefficients and feature noise:

$$\min_{\hat{\mathbf{w}}} \|\hat{\mathbf{w}}\|_1, \quad s.t. \quad \mathbf{x} = \mathbf{B}\hat{\mathbf{w}}, \quad (2)$$

where $\mathbf{B} = [\mathbf{D}; \mathbf{I}] \in \mathbb{R}^{d \times (n+d)}$ and $\hat{\mathbf{w}} = [\mathbf{w}^T; \xi^T]^T$. This optimization problem is convex and can be transformed into a general linear programming problem. There exists a globally optimal solution, and the optimization can be solved efficiently using many available ℓ_1 -norm optimization toolboxes like [ℓ_1 MAGIC]. Note that the ℓ_1 -norm optimization toolbox may convert the original constrained optimization problem into an unconstrained one, with an extra regularization coefficient which can be tuned for optimality in practice but essentially does not exist in original problem formulation.

Let $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_l, \mathbf{x}_{l+1}, \dots, \mathbf{x}_N\}$ be the set of feature vectors for the N images in the dataset, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the i -th sample in the dataset, and $\mathcal{G} = \{\mathcal{X}, \mathbf{W} = \{w_{ij}\}\}$ be the sparse graph with the samples in set \mathcal{X} as graph vertices and \mathbf{W} as the edge weight matrix. The construction of the k NN-sparse graph can be summarized as follows:

- (1) For each sample \mathbf{x}_i , search its k nearest neighbors $\mathcal{N}(\mathbf{x}_i)$. Here approximate method [Mount and Arya 1997] can be applied to accelerate the process.
- (2) Form the matrix \mathbf{B}_i with all samples $\mathbf{x}_{i_p} \in \mathcal{N}(\mathbf{x}_i)$: $\mathbf{B}_i = [\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_k}, \mathbf{I}] \in \mathbb{R}^{d \times (d+k)}$, where $p \in \{1, 2, \dots, k\}$ and $i_p \in \{1, 2, \dots, N\}$. Then the vector of the reconstruction coefficients for \mathbf{x}_i can be obtained by solving the following ℓ_1 -norm minimization problem:

$$\min_{\mathbf{w}_i} \|\mathbf{w}_i\|_1, \quad s.t. \quad \mathbf{x}_i = \mathbf{B}_i \mathbf{w}_i, \quad (3)$$

where $\mathbf{w}_i \in \mathbb{R}^{d+k}$. We call this one-vs- k NN sparse reconstruction. Note that if we set $\mathcal{N}(\mathbf{x}_i) = \{\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_N\}$, then it becomes the one-vs-all sparse reconstruction and $k = N - 1$.

- (3) Set the edge weight w_{ij} from the sample \mathbf{x}_j to the sample \mathbf{x}_i as:

$$w_{ij} = \begin{cases} \mathbf{w}_i(p), & \text{if } \mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i) \text{ and } j = i_p \\ 0, & \text{if } \mathbf{x}_j \notin \mathcal{N}(\mathbf{x}_i) \end{cases} \quad (4)$$

where $i, j \in \{1, 2, \dots, N\}$, and $\mathbf{w}_i(p)$ denotes the p -th element of vector \mathbf{w}_i .

3.2 Semi-Supervised Inference

Here we re-order the samples in set \mathcal{X} and have $\mathcal{X} = \mathcal{L} \cup \mathcal{U}$, where $\mathcal{L} = \{\mathbf{x}_1, \dots, \mathbf{x}_l\}$ contains the first l samples labeled as $y_i \in \{1, 0\}$ for every concept, and $\mathcal{U} = \{\mathbf{x}_{l+1}, \dots, \mathbf{x}_N\}$ contains the unlabeled ones. The label “1” indicates that the sample is relevant to a certain concept and “0” otherwise. It is well known that directly optimizing the binary label 1/0 is an NP-hard problem. Thus these labels are usually relaxed to be of real values. Each real-value label can be seen as the relevance score of the sample to a certain concept. As the objective of the semantic label inference is to rank the unlabeled samples according to their relevance values to each concept, so the real-value scores are naturally suitable. Denote the vector

of the predicted labels of all samples as \mathbf{f} , which can be split into two blocks as:

$$\mathbf{f} = \begin{bmatrix} \mathbf{f}_{\mathcal{L}} \\ \mathbf{f}_{\mathcal{U}} \end{bmatrix}. \quad (5)$$

Similar to the assumption in linear neighborhood propagation algorithm [Wang and Zhang 2008], we assume that the label of each sample can be reconstructed from the other samples, while the reconstruction coefficients are the same as those for the sparse reconstruction of sample vectors. Thus the linear reconstruction coefficients in the constructed sparse matrix can be used to predict the labels of the unlabeled samples. This prediction is based on the intuition that the weight w_{ij} reflects the likelihood for sample \mathbf{x}_i to have the same label as sample \mathbf{x}_j .

Based on the label reconstruction assumption, we can infer the labels of the unlabeled samples by minimizing the label reconstruction error as follows:

$$\min_{\mathbf{f}} \sum_{i=1}^N \|f_i - \sum_{j \neq i} w_{ij} f_j\|^2, \text{ s.t. } f_i = y_i, \text{ if } \mathbf{x}_i \in \mathcal{L}. \quad (6)$$

This formulation can be represented in matrix form as:

$$\min_{\mathbf{f}} [(\mathbf{I} - \mathbf{W})\mathbf{f}]^T [(\mathbf{I} - \mathbf{W})\mathbf{f}], \text{ s.t. } \mathbf{f}_{\mathcal{L}} = \mathbf{y}, \quad (7)$$

where \mathbf{y} is the label vector for the first l samples. Let $\mathbf{C} = (\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$ and differentiate the right side of Eqn.(7) with respect to \mathbf{f} , we obtain:

$$(\mathbf{C} + \mathbf{C}^T)\mathbf{f} = \mathbf{M}\mathbf{f} = \mathbf{0}, \quad (8)$$

where $\mathbf{M} = \mathbf{C} + \mathbf{C}^T$ is a symmetric matrix.

By splitting the matrix \mathbf{M} after the l -th row and l -th column, we have:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{\mathcal{L}\mathcal{L}} & \mathbf{M}_{\mathcal{L}\mathcal{U}} \\ \mathbf{M}_{\mathcal{U}\mathcal{L}} & \mathbf{M}_{\mathcal{U}\mathcal{U}} \end{bmatrix}, \quad (9)$$

and Eqn.(8) can then be rewritten as:

$$\begin{cases} \mathbf{M}_{\mathcal{L}\mathcal{L}}\mathbf{f}_{\mathcal{L}} + \mathbf{M}_{\mathcal{L}\mathcal{U}}\mathbf{f}_{\mathcal{U}} = \mathbf{0}, \\ \mathbf{M}_{\mathcal{U}\mathcal{L}}\mathbf{f}_{\mathcal{L}} + \mathbf{M}_{\mathcal{U}\mathcal{U}}\mathbf{f}_{\mathcal{U}} = \mathbf{0}. \end{cases} \quad (10)$$

By solving the second equation, we can obtain the label vector for the unlabeled samples:

$$\mathbf{f}_{\mathcal{U}}^* = -\mathbf{M}_{\mathcal{U}\mathcal{U}}^{-1}\mathbf{M}_{\mathcal{U}\mathcal{L}}\mathbf{y}. \quad (11)$$

Typically the matrix $\mathbf{M}_{\mathcal{U}\mathcal{U}}$ is very large for image annotation and retrieval tasks. It is often computationally prohibitive to calculate its inverse directly. Some conventional methods enforce the non-negative constraints for the reconstruction coefficients and use the iterative label propagation method to solve this problem. However, generally enforcing a non-negative constraint is not reasonable since some samples may have negative contributions to some other samples. Fortunately, the second equation of (10) can be reformulated as:

$$\mathbf{M}_{\mathcal{U}\mathcal{U}}\mathbf{f}_{\mathcal{U}} = -\mathbf{M}_{\mathcal{U}\mathcal{L}}\mathbf{y}. \quad (12)$$

The generalized minimum residual method (usually abbreviated as GMRES) [Saad and Schultz 1986] can be used to iteratively solve this large-scale sparse system of linear equations effectively and efficiently. The GMRES method approximates the solution by a vector in a Krylov subspace [Saad 2003] with minimal residue.

3.3 Handling of Noisy Training Labels

As aforementioned, the associated tags are often noisy for those community-contributed images. A quantitative analysis of noise in the associated tags for Flickr images can be found in [Chua et al. 2009]. Thus it is necessary to recover these noisy tags for achieving satisfactory image annotation performance if we use them as training labels.

To handle the noise in the training labels, we cannot assume that the training labels are fixed during the inference process as in Eqn.(6). The noisy training labels should be refined during the label inference step. However, they should be still consistent with the original labels to some extent. To handle these noisy labels, we propose to infer the labels of those unlabeled samples by adding two regularization terms:

$$\min_{\mathbf{f}} \{ \|\mathbf{f} - \mathbf{W}\mathbf{f}\|^2 + \lambda_1 \|\mathbf{f}_{\mathcal{L}} - \hat{\mathbf{f}}_{\mathcal{L}}\|^2 + \lambda_2 \|\hat{\mathbf{f}}_{\mathcal{L}} - \mathbf{y}\|_1 \}, \quad (13)$$

where $\mathbf{f}_{\mathcal{L}}$ encodes the training samples' labels that are propagatable on the sparse graph, and $\hat{\mathbf{f}}_{\mathcal{L}}$ denotes the ideal label vector of the training samples. The first term of this formula is the same as in Eqn.(6). The second term enforces the ideal labels of the training samples to be consistent with the labels propagatable on the derived sparse graph. This term essentially measures the quantity of the content-to-label noise. The third term is an ℓ_1 -norm, which measures the sparsity of the tag noise, and the minimization of which constrains that only a few elements are different between the ideal labels and the original labels, since generally only a limited number of labels are noisy. The intuitive explanation for the regularization is that the training labels should be consistent to both the original tags and the inferred labels of the training samples.

This problem can be solved in three steps:

- (1) Set the original label vector as the initial estimation of ideal label vector, that is, set $\hat{\mathbf{f}}_{\mathcal{L}} = \mathbf{y}$, and then solve

$$\min_{\mathbf{f}} \{ \|\mathbf{f} - \mathbf{W}\mathbf{f}\|^2 + \lambda_1 \|\mathbf{f}_{\mathcal{L}} - \hat{\mathbf{f}}_{\mathcal{L}}\|^2 \}. \quad (14)$$

It can be solved similarly as in Section 3.2, and we can obtain a refined $\mathbf{f}_{\mathcal{L}}$.

- (2) Fix $\mathbf{f}_{\mathcal{L}}$ and solve

$$\min_{\hat{\mathbf{f}}_{\mathcal{L}}} \{ \|\mathbf{f}_{\mathcal{L}} - \hat{\mathbf{f}}_{\mathcal{L}}\|^2 + \frac{\lambda_2}{\lambda_1} \|\hat{\mathbf{f}}_{\mathcal{L}} - \mathbf{y}\|_1 \}. \quad (15)$$

It is an ℓ_1 -norm minimization problem, which can be solved using the toolboxes such as [ℓ_1 MAGIC].

- (3) Use the obtained $\hat{\mathbf{f}}_{\mathcal{L}}$ to replace the \mathbf{y} in Eqn.(12), and we can solve the sparse system of linear equations to infer the labels of the unlabeled samples.

Note that the first and second steps can be iterated several times for more robust removal of tag noise. Our experiments show that generally one iteration is enough to achieve a sufficiently stable solution.

Table I. Abbreviations of the Method Names.

Abbreviation	Full name of the method
SVM	Support Vector Machine [Chang and Lin 2001]
k NN	k Nearest Neighbors [Duda et al. 2000]
LNP	Linear Neighborhood Propagation [Wang and Zhang 2008]
S-Recon	Sparse Label Reconstruction [Wright et al. 2009]
SGSSL	Sparse Graph-based Semi-Supervised Learning [Tang et al. 2009]
k NN-SGSSL	k NN-Sparse Graph-based Semi-Supervised Learning
A k NN-SGSSL	Approximate k NN-SGSSL
A k NN-SGSSL _{dn}	Approximate k NN-SGSSL with noisy label handling

4. EXPERIMENTS

To evaluate the performance of the proposed k NN-sparse graph-based semi-supervised learning method, we conducted extensive experiments on a real-world community-contributed image dataset along with their associated tags. In Table I, we abbreviate the names of all the compared methods in the experiments.

4.1 Dataset

The dataset we used in all the experiments is a lite version of the NUS-WIDE database [Chua et al. 2009]. This dataset includes 55,615 images and their associated tags, which are crawled from Flickr. We use half of these images (*i.e.* 27,807 images) for training by *using their associated noisy tags as training labels*, and the rest (*i.e.* 27,808 images) for testing by ignoring the crawled tags associated to the test images. The data separation is the same as in [Chua et al. 2009]. Annotation performances are evaluated based on the 81 concepts defined in NUS-WIDE, where the ground-truth of these concepts for all images are provided for evaluation.

The low-level features we used here include the 64-D color histogram and 73-D edge direction histogram. We combine these two kinds of features directly by merging the two feature vectors of each sample. We subtract every element of each dimension of features by the mean of all elements in this dimension, and then divide by three times the standard variation of all elements in this dimension to normalize the features.

For each concept, the test images are ranked according to the probability that the images are relevant. The performance is measured via non-interpolated Average Precision (AP), a standard metric used for image retrieval [TREC]. We average the APs over all the evaluated 81 concepts to create the Mean Average Precision (MAP), which is an overall performance measure. All approaches in the experiments are executed on a PC with Intel 3.0GHz CPU and 16G memory.

4.2 Comparisons Among SGSSL and Other Traditional Learning Methods

To demonstrate the effectiveness of the sparse graph-based semi-supervised learning approach (SGSSL), we compare its performance with the following four methods for image annotation: SVM, k NN, LNP, and S-Recon. LNP is one of the state-of-art graph-based semi-supervised learning methods. Different from constructing the graph by sparse reconstruction that is achieved by ℓ_1 minimization, it constructs the graph by minimizing the ℓ_2 -norm of reconstruction error $\varepsilon_i = \|\mathbf{x}_i - \sum_{j:\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} w_{ij} \mathbf{x}_j\|^2$. S-Recon can be regarded as a supervised version

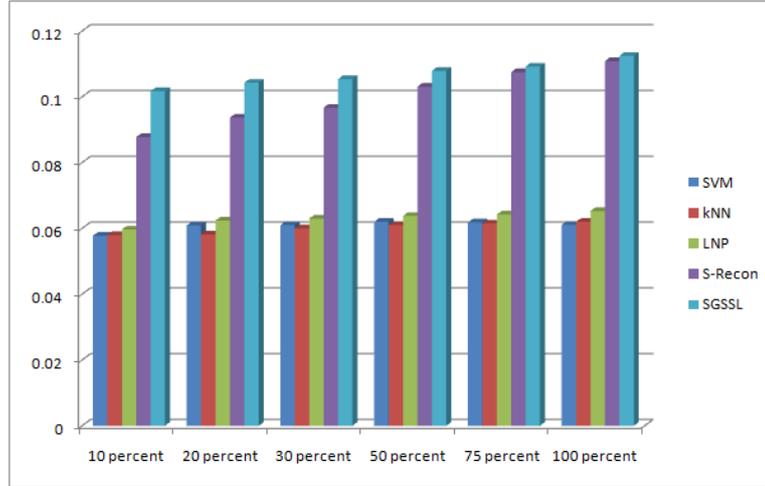


Fig. 3. The comparison of MAPs of the five methods using different proportions of training data.

Table II. Comparison of execution times for the five methods evaluated.

Method	Executing Time (minutes)
SVM	2400
kNN	80
LNP	900
S-Recon	130
SGSSL	300

of the approaches in this paper. It reconstructs the labels of the given samples directly using the same sparse coefficients that reconstruct the samples: if the sparse coefficient vector of reconstructing the given sample \mathbf{x}_i is \mathbf{w}_i (refer to formula (3)), the label of \mathbf{x}_i can be predicted as: $f_i = \mathbf{w}_i * \mathbf{f}_L$.

We conducted six groups of experiments to compare the performances of these five methods by using different proportions of training set: 10 percent, 20 percent, 30 percent, 50 percent, 75 percent, and 100 percent (*i.e.*, the entire training set). When we use partial training data for inference, the samples are randomly sampled from the entire training set according to the given proportions. All evaluations were conducted on the same test dataset. The MAPs of these five methods using different proportions of training data are illustrated in Fig. 3. From these results, we can observe that:

- (1) For all cases SGSSL outperforms the first three methods significantly.
- (2) SGSSL always outperforms S-Recon when using different proportions of training data, and the performance of SGSSL is much better than that of S-Recon when the size of training data is small. For example, when using 10 percent of training data, SGSSL obtains a MAP of 0.102, which has an improvement of 15.8% over S-Recon. When the size of training data increases, the performance of S-Recon becomes comparable to SGSSL.
- (3) The performances of both SGSSL and S-Recon increase accompanied with the

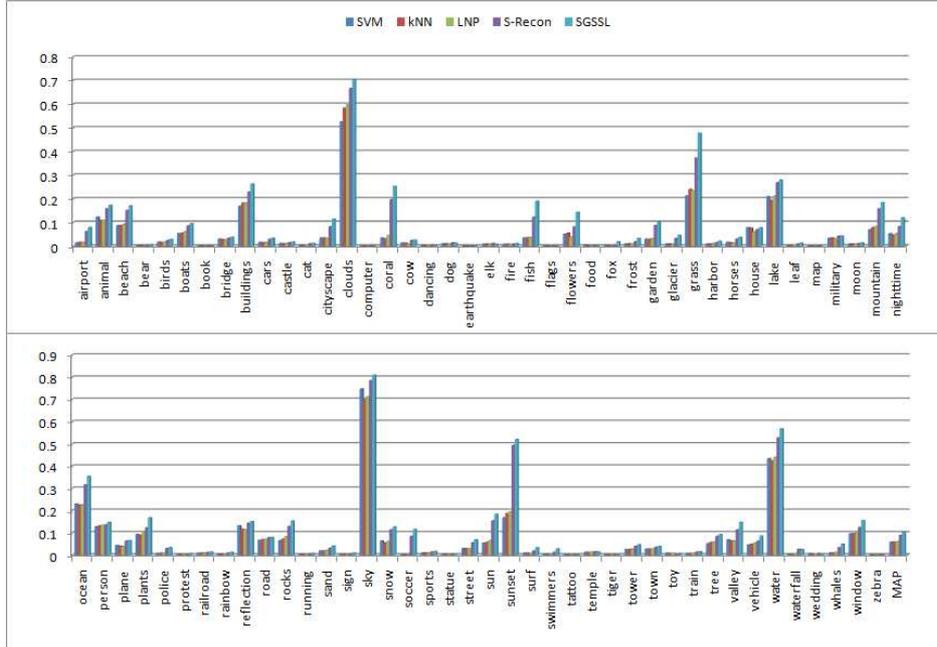


Fig. 4. The comparison of APs for the 81 concepts using the five methods with 10 percent of training data.

increment of the training data size, while the performance of SGSSL increases slower than that of S-Recon. The reason is obvious: semi-supervised learning utilizes the distribution of unlabeled data to boost the performance, when the size of training data is large enough, there is no need to utilize this distribution information.

Due to the limitation of space, we only provide the comparison of APs for different concepts using 10 percent of training data as shown in Fig. 4. From these results, we can observe that SGSSL outperforms the other four methods for almost all the 81 concepts.

The approximate execution times of these five methods for annotating the 81 concepts with 10 percent of training data are given in Table II. Here $k = 300$ in k NN search. Although SGSSL is much slower than k NN, it is significantly more effective. Also it is slower than S-Recon when using 10 percent of training data but with significant improvement in annotation accuracy. However, when the size of training data increases, the annotation accuracy of S-Recon will increase to be comparable with SGSSL, but its execution time will also increase to be similar to that of SGSSL.

4.3 Comparisons Among SGSSL, k NN-SGSSL and Approximate k NN-SGSSL

As the one-vs-all sparse reconstruction is very time-consuming, we reconstruct each sample from its k nearest neighbors to form the k NN-sparse graph to reduce the computational complexity. Meanwhile, an approximate method [Mount and Arya

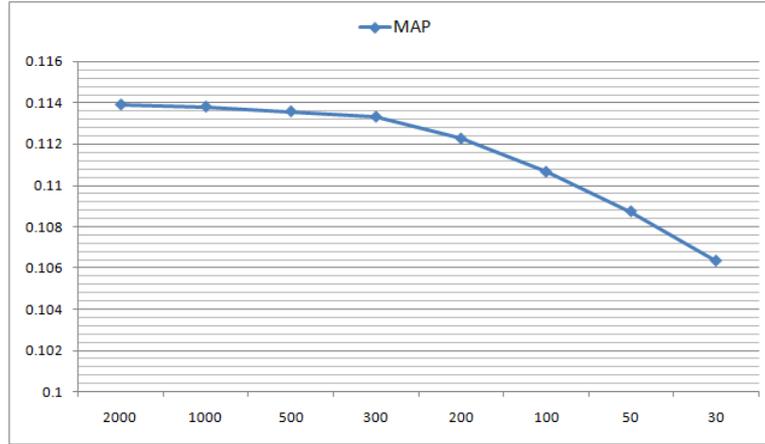


Fig. 5. The MAPs of k NN-SGSSL with different k .

Table III. Executing time (minutes) of each step in the SGSSL, k NN-SGSSL, Ak NN-SGSSL and Ak NN-SGSSL_dn using the entire training set.

	SGSSL	k NN-SGSSL	Ak NN-SGSSL	Ak NN-SGSSL_dn
k NN Search	—	250	40	40
Sparse Graph Construction	630	5	5	5
Prediction	10	10	10	15
Total executing time	640	265	55	60

1997] is also adopted to accelerate the k NN search. In this section, we evaluate the performances of label propagation on the k NN-sparse graph and approximate k NN-sparse graph. All experiments use the entire training set for label propagation.

Fig. 5 compares the MAPs of k NN-SGSSL with different k . We can see that the performance does not change too much when k exceeds 300, hence we set the value of k to 300 for the rest of experiments. In Fig. 6, we compare the APs obtained by SGSSL, k NN-SGSSL, and Ak NN-SGSSL, with $k=300$. From these results, we can see that the effectiveness of these three methods are similar for image annotation. The MAP obtained by k NN-SGSSL is 0.1133, which is even a bit better than the 0.1123 obtained by SGSSL. This is because that the constructed k NN-sparse graph utilizes both the visual distance and sparse reconstruction, thus it is better than using only the sparse reconstruction. The Ak NN-SGSSL obtains a MAP of 0.1124, which is similar to that of k NN-SGSSL. That is to say, the approximate k NN nearly does not affect the effectiveness compared to the accurate k NN.

Table III illustrates the executing time of each step in the SGSSL, k NN-SGSSL, Ak NN-SGSSL and Ak NN-SGSSL_dn. We can see that k NN-SGSSL is much more efficient than SGSSL, and Ak NN-SGSSL further reduces the computational complexity significantly.

4.4 Approximate k NN-SGSSL with Noisy Training Label Refinement

To evaluate the effectiveness of the training label refinement strategy, we compare the performances of approximate k NN-SGSSL with and without training la-

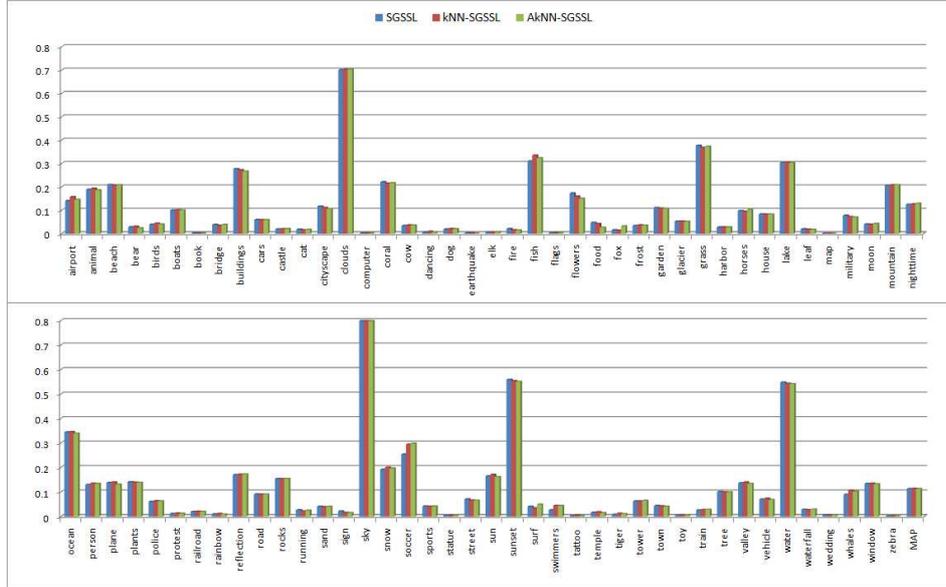


Fig. 6. The APs obtained by SGSSL, k NN-SGSSL and Ak NN-SGSSL.

bel refinement. We empirically set the parameters for the handling of label noise: $\lambda_1 = 100$ for Eqn.(14) and $\frac{\lambda_2}{\lambda_1} = \frac{1}{5}$ for Eqn.(15). Here in Eqn.(14), the λ_1 constrains that the ideal labels for the labeled samples should change but should not change too much compared to the initial labels. Thus it should be much larger than 1. And in Eqn.(15), the $\frac{\lambda_2}{\lambda_1}$ should be able to control the balance between the two regularization terms.

Fig. 7 compares the APs obtained by Ak NN-SGSSL and Ak NN-SGSSL_{dn}. We can see that after handling the noise in the training labels, Ak NN-SGSSL_{dn} outperforms Ak NN-SGSSL significantly for almost all the 81 concepts. Ak NN-SGSSL_{dn} achieves an MAP of 0.1428, which has an improvement of 27.1% over Ak NN-SGSSL. It indicates that the training label refinement step is critical and valuable for image annotation with noisy training labels.

The executing time of Ak NN-SGSSL_{dn} is comparable to that of Ak NN-SGSSL (see Table III). Thus we can see that performing noisy training label refinement will not bring too much additional computational cost but will bring much extra benefits in effectiveness.

5. CONCLUSIONS AND FUTURE WORK

In this work, we exploited the problem of annotating images by label propagation over community-contributed images and their associated noisy tags. A novel k NN-sparse graph-based semi-supervised learning approach was proposed to improve the annotation performance by handling the links among the semantically-unrelated samples. Meanwhile, the approximate k NN search is applied to ensure the efficiency. In addition, an effective training label refinement strategy is proposed into the graph-based learning framework to reduce the effects of noise in

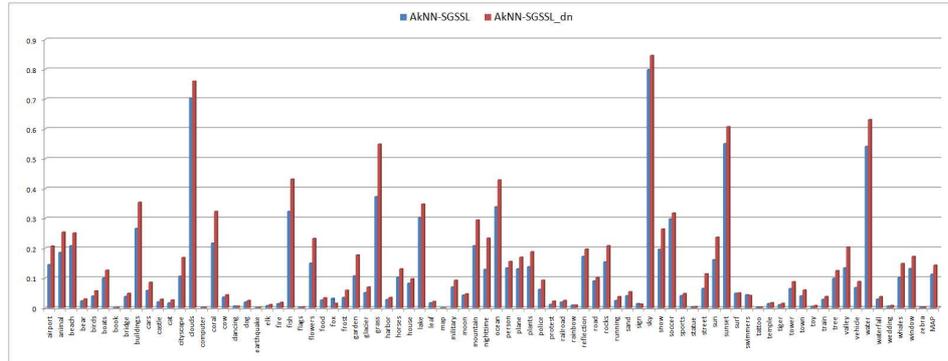


Fig. 7. The APs obtained by AkNN-SGSSL and AkNN-SGSSL_{dn}.

the tags. Extensive experiments conducted on the NUS-WIDE-Lite dataset have demonstrated the effectiveness and efficiency of the proposed approach. From the experimental results, we can see that a key factor, which affects the performance of image annotation with the tags as training labels, is the noise in tags. Actually, for image annotation, we may not need to correct all the noisy tags. Instead we can collect the correct image-label pairs as much as possible for training. Thus our future work will focus on how to construct an effective training set from the community-contributed images and tags.

6. ACKNOWLEDGEMENT

This work is partially supported by NRF/IDM Program of Singapore, under Research Grants NRF2007IDM-IDM002-047 and NRF2008IDM-IDM004-029.

REFERENCES

- BELKIN, M. AND NIYOGI, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*.
- CHANG, C.-C. AND LIN, C.-J. 2001. LIBSVM: a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- CHAPELLE, O., ZIEN, A., AND SCHOLKOPF, B. 2006. *Semi-supervised Learning*. MIT Press.
- CHUA, T.-S., TANG, J., HONG, R., LI, H., LUO, Z., AND ZHENG, Y.-T. 2009. NUS-WIDE: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval*. Santorini, Greece.
- DONOHO, D. L. 2006. For most large underdetermined systems of linear equations the minimal ℓ_1 -norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics* 59, 6, 797–829.
- DUDA, R., STORK, D., AND HART, P. 2000. *Pattern Classification*. JOHN WILEY.
- ℓ_1 MAGIC. <http://www.acm.caltech.edu/l1magic/>.
- FERGUS, R., FEI-FEI, L., PERONA, P., AND ZISSERMAN, A. 2005. Learning object categories from google’s image search. In *IEEE International Conference on Computer Vision*.
- GOH, K.-S., CHANG, E. Y., AND LAI, W.-C. 2004. Multimodal concept-dependent active learning for image retrieval. In *Proc. of the 12th annual ACM international conference on Multimedia*. 564–571.
- HE, J., LI, M., ZHANG, H.-J., TONG, H., AND ZHANG, C. 2004. Manifold-ranking based image retrieval. In *ACM Multimedia*.

- LI, X., CHEN, L., ZHANG, L., LIN, F., AND MA, W.-Y. 2006. Image annotation by large-scale content-based image retrieval. In *ACM Multimedia*.
- MOUNT, D. AND ARYA, S. 1997. Ann: A library for approximate nearest neighbor searching. In *CGC 2nd Annual Fall Workshop on Computational Geometry*.
- RAO, R., OLSHAUSEN, B., AND LEWICKI, M. 2002. *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press.
- ROWEIS, S. T. AND SAUL, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326.
- SAAD, Y. 2003. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Second Edition.
- SAAD, Y. AND SCHULTZ, M. 1986. Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing* 7, 856–869.
- SUN, Y., SHIMADA, S., TANIGUCHI, Y., AND KOJIMA, A. 2008. A novel region-based approach to visual concept modeling using web images. In *Proceeding of the 16th ACM International Conference on Multimedia*. Canada.
- TANG, J., HUA, X.-S., QI, G.-J., WANG, M., MEI, T., AND WU, X. 2007. Structure-sensitive manifold ranking for video concept detection. In *ACM Multimedia*. Augsburg, Germany.
- TANG, J., HUA, X.-S., SONG, Y., QI, G.-J., AND WU, X. 2008. Video annotation based on kernel linear neighborhood propagation. *IEEE Transaction on Multimedia* 10, 4.
- TANG, J., YAN, S., HONG, R., QI, G.-J., AND CHUA, T.-S. 2009. Inferring semantic concepts from community-contributed images and noisy tags. In *Proceedings of the seventeen ACM international conference on Multimedia*. 223–232.
- TORRALBA, A., FERGUS, R., AND FREEMAN, W. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 30, 11.
- TREC. Trec-10 proceedings appendix on common evaluation measures. <http://trec.nist.gov/pubs/trec10/appendices/measures.pdf>.
- WANG, C., JING, F., ZHANG, L., AND ZHANG, H.-J. 2006. Image annotation refinement using random walk with restarts. In *Proc. ACM Multimedia*.
- WANG, F. AND ZHANG, C. 2008. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering* 20, 1, 55–67.
- WANG, X.-J., ZHANG, L., JING, F., AND MA, W.-Y. 2006. Annosearch: Image auto-annotation by search. In *IEEE Conference on Computer Vision and Pattern Recognition*. New York, USA.
- WANG, X.-J., ZHANG, L., LI, X., AND MA, W.-Y. 2008. Annotating images by mining image search results. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 30, 11, 1919–1932.
- WRIGHT, J., YANG, A., GANESH, A., SASTRY, S., AND MA, Y. 2009. Robust face recognition via sparse representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 31, 2 (Feb.), 210–227.
- ZHOU, D., BOUSQUET, O., LAL, T. N., WESTON, J., AND SCHOLKOPF, B. 2003. Learning with local and global consistency. In *Proc. 17-th Annual Conference on Neural Information Processing Systems*.
- ZHU, X. 2005. *Semi-Supervised Learning with Graphs*. PhD Thesis, CMU.
- ZHU, X., GHAHRAMANI, Z., AND LAFFERTY, J. 2003. Semi-supervised learning using gaussian fields and harmonic function. In *Proc. 20-th International Conference on Machine Learning*.