Image Tagging by Semantic Neighbor Learning Using User-Contributed Social Image Datasets

Feng Tian  
*the School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China.*

Xukun Shen  
*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China.*

Xianmei Liu  
*the School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China.*

Maojun Cao  
*the School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China.*

Follow this and additional works at: [https://tsinghuauniversitypress.researchcommons.org/tsinghua-science-and-technology](https://tsinghuauniversitypress.researchcommons.org/tsinghua-science-and-technology)

Part of the Computer Sciences Commons, and the Electrical and Computer Engineering Commons

**Recommended Citation**  
Image Tagging by Semantic Neighbor Learning Using User-Contributed Social Image Datasets

Feng Tian*, Xukun Shen, Xianmei Liu, and Maojun Cao

Abstract: The explosive increase in the number of images on the Internet has brought with it the great challenge of how to effectively index, retrieve, and organize these resources. Assigning proper tags to the visual content is key to the success of many applications such as image retrieval and content mining. Although recent years have witnessed many advances in image tagging, these methods have limitations when applied to high-quality and large-scale training data that are expensive to obtain. In this paper, we propose a novel semantic neighbor learning method based on user-contributed social image datasets that can be acquired from the Web's inexhaustible social image content. In contrast to existing image tagging approaches that rely on high-quality image-tag supervision, we acquire weak supervision of our neighbor learning method by progressive neighborhood retrieval from noisy and diverse user-contributed image collections. The retrieved neighbor images are not only visually alike and partially correlated but also semantically related. We offer a step-by-step and easy-to-use implementation for the proposed method. Extensive experimentation on several datasets demonstrates that the performance of the proposed method significantly outperforms others.

Key words: image tag; social image tagging; user-contributed datasets; semantic neighbor learning

1 Introduction

Image tagging is the process of assigning tags to images to describe their content[1]. With these tags, users can easily retrieve and manage images. Since manual image tagging is time-consuming and labor intensive, automatic image tagging has attracted great interest. Despite significant progress in recent years, the effectiveness of existing image tagging methods is heavily dependent on high-quality datasets, which requires intense manual work. Recently, image sharing communities have evolved to the degree that it is now easier to obtain a large number of images with associated tags[2–7].

Given that a number of tagged images are publicly accessible, is it now possible for us to use socially tagged images as training data to develop stronger models? This is an interesting and difficult question. The image database known as ImageNet[8] provides the answer. This database contains positive images in over 20 000 classes that are organized according to WordNet. However, these positive images tend to be biased by the search results. For example, in ImageNet, the set of images tagged by the word “vehicle” consists primarily of cars, despite the fact that trucks, watercraft, and aircraft also belong to the vehicle class[9]. We note that the diverse visual appearance of social images is very important for developing effective models, and social tagging results generate more diverse training images. We believe that the degree of learning achieved based on social training data can be improved with a better image tagging method.
Several methods for image tagging in social frameworks have been proposed\cite{10-19}. Generative models like Statistical Machine Learning (SML) treat image tagging as a classification problem, with each class comprising one group of images labeled with one tag\cite{10}. The mixture density is estimated for each image and the mixtures of each class are then pooled into the mixture of the corresponding tag. These tags are then assigned to images by computing the probability of minimum error tagging. This method is based on the assumption that the training image tags are accurate, objective, and uniformly distributed. However, this assumption is generally invalid in social training sets in which noise occurs due to incompleteness and tag diversity.

To learn from data in which some samples are only partially tagged, a convex quadratic optimization method has been proposed that uses hinge loss to minimize the regularized empirical risk\cite{11}. However, this method is also based on the assumption that all of the given tags are correct. In Ref.\cite{12}, the authors proposed a hybrid model for utilizing partially tagged data to improve image segmentation performance and the results show that integrating semantic information and co-occurrences of image feature patterns into one framework is helpful in handling incomplete tags. Compared with the complex process of model learning, nearest-neighbor-based methods have become more popular recently as the amount of training data has rapidly increased. Among others, Joint Equal Contribution (JEC)\cite{13} is a simple yet effective method for social training sets, whereby nearest neighbors are determined by the average of several distances computed based on different visual features, and then the tags are transferred from neighbors to the given image. In contrast to JEC, in which visual neighbors are treated equally, TagProp\cite{14} predicts tags using neighbor voting plus distance metric learning. TagProp also promotes rare tags and penalizes frequent tags by training a logistic per-tag model. The effectiveness of this method has been verified on the Flickr dataset\cite{15}. In contrast to JEC, which treats all features equally, and Tagprop, which concatenates different features to form a global descriptor, a feature selection algorithm\cite{16} has been proposed that leverages the sparsity of features to determine image similarity, and then transfers the tags of the most similar images to the test image. The results show that using the sparsity of visual features can boost image tagging performance.

To deal with imprecise or fuzzy tags of social images, researchers have developed a tag cleansing algorithm\cite{17}. To handle incompletely tagged data, another group explored the use of group lasso regularizer for estimating the ranking error of assigned classes\cite{18}. However, again, this method is based on the assumption that the given sample tags are correct. In Ref.\cite{19}, image classifiers and tag classifiers are jointly trained and then agree upon the list of tags predicted for each image.

Although the above studies have made advances in image tagging, image tagging performance when using social training sets remains unsatisfactory due to the following unresolved challenges. First, social tags tend to be ambiguous, imprecise, and incomplete\cite{1}. Second, due to their various interests and motivations, users establish personalized and biased image tags, as illustrated in Fig. 1. For example, images tagged with “car” might be obtained when a user is driving a car rather than depicts the car itself. Existing research shows that most images (more than 50% in the Flickr database) have no tag at all, and only half of the tags are visually associated with the image. This also applies to benchmark datasets. Fortunately, image tags are provided by a large number of heterogeneous users. We believe that these tags are complementary. As such, it is important to take into account tag correctness and completeness. Moreover, due to the distribution bias of social tags, learning based on these data will yield unsatisfactory performance with respect to rare tags. For example, we found that with rare tags, JEC\cite{13} achieved 19% in terms of the F1 value. In contrast, it achieved 51% for most common tags.

Actually, the essence of effective image tagging lies in how original noisy tags are handled. Unfortunately, since the social tags of images are very diverse, noisy, and sparse, the direct use of social image datasets is ineffective and this issue differs significantly from

![Fig. 1 Image examples in user-contributed datasets (missing tags are highlighted in bold and content-unrelated tags are underlined).](image-url)
the problem of classical image tagging. To deal with ambiguous, imprecise, and incomplete tags in social training sets, we present a novel image-tagging approach based on Semantic Neighbor Learning (SNL). Figure 2 shows an overview of our semantic neighbor learning method for image-tagging.

First, to reduce the sparsity of the tags, we replenish tags by exploring original tags. Then, we generate a neighborhood from different semantic groups. Since both common and rare tags will occur in this neighborhood, we refer to it as a “semantically balanced neighborhood”. Next, we construct a more accurate neighborhood using metric learning with multiple tags and sparse reconstruction. Since samples in the neighborhood are visually alike, partially correlated, and semantically related, we call it a “semantically consistent neighborhood”. Finally, we apply semi-supervised tag inference in the “semantically consistent neighborhood”. To demonstrate the advantages of our proposed semantic neighbor learning method, we conducted extensive experiments on both classical and social image datasets extracted from Flickr.

The main contributions of this work can be summarized as follows:

(1) We propose a novel semantic neighbor learning method that employs progressive neighborhood retrieval from noisy and diverse user-contributed image collections. The set of associated tags in the neighborhood contains richer information to describe a corresponding image. In addition, the data distribution in the collection is well balanced. Furthermore, we remove most semantically unrelated neighbors by generating a “semantically consistent neighborhood”. The retrieved neighbors are not only visually alike but also more consistent with human perception. More importantly, the obtained neighbor set is more robust with respect to noisy elements. Our key concept and algorithm can easily be extended to other modalities (e.g., video tagging, 3-D model tagging, and audio tagging).

(2) We develop an efficient semi-supervised image tagging algorithm to select a small number of tags that are most probably semantically related and assign these optimal tags to the image. The proposed algorithm is an easy-to-use method that can avoid the propagation of noisy tags, and is compatible with other methods.

The remainder of this paper is organized as follows. In Section 2, we review related work. In Section 3, we describe in detail our proposed SNL model and tag inference method. We report our experimental settings and results in Section 4, and draw our conclusions and suggest future work in Section 5.

2 Related Work

In this section, we discuss a number of representative research efforts for image-tagging. Generally speaking, existing methods can be classified into two groups. The first group contains traditional image tagging methods that learn from high-quality image datasets and the second includes methods suitable for social image datasets.

2.1 Traditional image tagging methods

Traditional image tagging methods rely heavily on image tags as a means of supervision. By nature, these methods are topic or mixture models[20], developed on the basis of high-quality training sets. These models can be tag-specific or integral to all tags. The topic model is the representative method for integral modeling[21–25], in which each topic represents a distribution over
visual features and tags. In tag-specific modeling, classifiers are trained for each tag. Many approaches have been proposed for this purpose, including linear Support Vector Machine (SVM) classifiers\cite{26}, fast-intersection kernel SVM classifiers\cite{27}, and ensembles of SVM classifiers\cite{28}. All of the above models have demonstrated good performance on traditional datasets. However, these approaches typically perform less well with datasets that lack high-quality training data. The problem becomes much more severe when dealing with real world datasets.

2.2 Social image tagging methods

With large-scale images and tags, it is reasonable to expect the introduction of image tagging based on social datasets. Recently, researchers have found social datasets in which tag quality is low to be useful for learning image tags for methods including JEC\cite{13}, neighbor voting\cite{18} and its variants\cite{19–34}, TagProp\cite{14, 15}, group sparsity\cite{16}, and FastTag\cite{19}. These data-driven methods assume that similar images share similar tags and they work surprisingly well by working to retrieve similar neighbors. The authors of Ref. [13] found that equal contributions from different visual features perform on a par with Lasso. The authors of Ref. [29] presented a neighbor voting algorithm that estimates the relevance of a tag with respect to an image by counting the number of occurrences of the tag in the neighborhood of the image. TagProp\cite{14, 15} was proposed to further improve the performance of image-tagging, which employs neighbor voting plus distance metric learning. A probabilistic framework was also proposed in which the probability of using images in the neighborhood is defined based on rank or distance-based weights. A feature selection algorithm was proposed using the sparsity of features in Ref. [16]. All of these methods have confirmed that better performance can be obtained by the application of an appropriate neighbor selection mechanism. Chen et al.\cite{19} proposed the FastTag method, which simultaneously learns two classifiers on two sources, i.e., image and text, and predicts an agreed-upon list of tags for each image. In Ref. [35], the authors filtered noisy images for cross-domain semantic transfer.

Recent advances in deep-feature learning also offer a promising route, and Convolutional Neural Networks (CNN), among others, have exhibited power with respect to image tagging\cite{36–38}. The work in Ref. [36] shows that a large and deep CNN is capable of achieving high performance using purely supervised learning. In Ref. [37], the authors extracted CNN features from an image and word-embedded vectors to represent their associated tags and then utilize them to tag images. Although the use of deep learning-features has been shown to yield better performance, there has been limited research on how to further fine tune these features for social image datasets. One reason for this might be that most deep-learning approaches occur fully supervised settings and thus have limited application in domains for which it is expensive to obtain high-quality and large-scale training data.

Therefore, for these user-contributed datasets, it is difficult to directly use ambiguous, imprecise, and incomplete social tags. In contrast, our method focuses on learning semantic neighbors from noisy and diverse user-contributed image collections. Moreover, we propose an efficient image tagging algorithm for transferring tags from obtained neighbors that is robust with respect to noise.

3 Semantic Neighbor Learning

The basic premise of image tagging from a user-contributed dataset is to obtain similar neighbors from this noisy dataset. Given a large amount of social images and tags, it is difficult for existing models to identify neighbors in social training sets due to their extreme sparsity (i.e., image tags are incomplete and fifty percent of the images are not tagged), diversity (i.e., the tags are not uniformly distributed), and noise (i.e., the tags are unrelated with respect to describing the visual content). Therefore, directly retrieving neighbors by their visual and tag similarities is not effective. Here, we propose a neighborhood learning method for retrieving neighbors that considers both the tags in the neighbor set and the relatedness of the images. More specifically, the retrieved neighbors are not only visually alike, but also partially visually correlated and semantically related. For example, images of “there is an elephant in the park to take a bath” should relate not to images of “elephant” but also to “park” and “water”. Thus, tags of these neighbors can be collected to describe the given image. The major advantages of our method are as follows:

Completeness and Fairness: We effectively replenish tags by exploring original tags in the
dataset by transductive inference, thereby significantly reducing the sparsity of the image tags. Furthermore, we generate a “semantically balanced neighborhood” from different semantic groups. As a result, both common and rare tags will occur in the neighbor set, so the tags in the neighbor set are more informative.

**Relatedness:** Compared to traditional methods, our method learns the metrics between neighbors and by doing so, retrieves most of the neighbors in one subspace. Furthermore, we generate “semantically consistent neighborhoods” by sparse reconstruction in the subspace, to remove most of the semantically unrelated neighbors. As a result, samples in the neighbor set are not only visually alike but also partially correlated and semantically related.

**Robustness:** We apply semi-supervised tag inference based on the learned similarity between neighbors, thereby involving only a small number of the most probably semantically related samples which are robust with respect to noise.

Table 1 lists the major notations used in this method.

### 3.1 Semantically balanced neighborhood construction

Considering the fact that many tags are omitted by users, especially rare tags, that are extracted from user-contributed tags, we propose a tag replenishment strategy to solve the problem of tag sparsity. In this strategy, we introduce the regularization of the strategy to solve the problem of tag sparsity. In other words, we define \( T = [t_1, \cdots, t_l]^T \) as the original tag indicator matrix and \( Y = [y_1, \cdots, y_l]^T \) as the replenished tag indicator matrix. We construct a \( \delta \)-nearest neighbor graph in which we set the edge weight to be \( w_{ij} = e^{-\frac{\|x_i-x_j\|^2}{2\sigma^2}} \) and the parameter \( \sigma \) is defined as the median of the distances over all image pairs. Therefore, the problem of tag replenishment is formulated as an optimization problem as follows:

\[
\min \left\{ \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} u_{ij} (t_{ij} - y_{ij})^2 + \right. \\
\left. \frac{1}{2} \lambda \sum_{i=1}^{l} \sum_{j=1}^{l} w_{ij} \left\| \frac{y_i}{\sqrt{d_i}} - \frac{y_j}{\sqrt{d_j}} \right\|^2 \right\}
\]

(1)

where \( \lambda \) is the regularization parameter, which penalizes disagreement between original and replenished tags, and the first regularizer facilitates the assignment of similar tags to similar images.

\[
d_i = \sum_{j=1}^{l} w_{ij}, \quad u_{ij} \text{ is the relevance between sample } x_i \text{ and the } j-th \text{ tag where } u_{ij} = 1 \text{ if } t_{ij} = 1 \text{, and } \tau \text{ otherwise } (0 < \tau < 1). \text{ The second regularizer can be rewritten as follows:}
\]

\[
\text{tr}(Y^T(I - V^{1/2} W V^{1/2})Y)
\]

(2)

where \( W = \{w_{ij}\} \), and \( V = \text{diag}([d_1, \cdots, d_l]) \). Given Formula (1), we take the derivative of it with respect to \( Y \) and have the following:

\[
U \odot [Y - T] + \lambda SY
\]

(3)

where \( U = \{u_{ij}\} \) and \( S = I - D^{1/2} W D^{1/2} \). By setting Formula (3) to zero, we can derive the following:

\[
(\text{diag}(U_{ij}) + \lambda S)Y_{ij} = Z_{ij}
\]

(4)

where \( Z_{ij} = U_{ij} \cdots T_{ij} \), and \( U_{ij} = [u_{1j}, \cdots, u_{lj}]^T \). We can easily derive the approximate optimal solution of Eq. (4) by the least squares.

By this replenishment, training images have more informative tags. Considering that rare tags will yield poor performance, we boost the frequency of rare tags and decrease the frequency of common tags by constructing semantic group \( L_i \subseteq L, \forall i \in \{1, 2, \cdots, q\} \), which is a subset that contains all images tagged with the \( i \)-th tag. Given an image \( x \), from each group we retrieve \( \delta \) images that are most visually similar to \( x \). All the neighbors make up the “semantically balanced neighborhood”, which we denote as \( \text{BN}(x) = \{L_{x,1} \cup \cdots \cup L_{x,q}\} \). In this way, both common and rare tags occur in the neighbor.
set, which makes the tags in the neighborhood more informative. Figure 3 shows a number of examples retrieved from “semantically balanced neighborhood”.

3.2 Semantically consistent neighborhood construction

By generating a “semantically balanced neighborhood”, both common and rare tags occur in the neighbor set, thus making the tags in the neighbor set more informative. However, some neighbors are not meaningful in describing given images, as illustrated in Fig. 4. With respect to the image in Fig. 4a, only the images in Figs. 4b and 4c are partially correlated with it, so they are more informative. The other neighbors, such as those shown in Figs. 4d and 4e are not semantically relevant to that in Fig. 4a. Considering that the effectiveness of image tagging depends on the similar neighbors used, we assume there to be an optimal neighbor set within which neighbors are not only visually alike but also partially visually correlated and semantically related. Neural scientists have found the human vision system to seek a sparse representation for an image using a few visual words. This fact motivated us to retrieve partially correlated neighbors in a neighbor set by the sparse reconstruction of the samples, thereby making these neighbors members of a “semantically consistent neighborhood”.

Fig. 3 Neighbor images retrieved from a “semantically balanced neighbor set” and from a JEC [7] neighborhood. For a given image (first column) along with its ground-truth tags, there are many rare tags in the neighbor set (the first row). In contrast, apart from “lake”, the neighbor set obtained by JEC [7] (second row) contains only frequent tags.

Fig. 4 Semantically balanced neighborhood of a given image.

Our rationale to focus on improving the effectiveness of neighborhood construction is as follows: (1) Semantically unrelated neighbors can be removed, thereby the effectiveness of image-tagging can be improved. (2) Each image has links only to a small number of semantically related images, thereby the efficiency of image-tagging can be improved.

To guarantee the physical meaning of sparse representation in the neighborhood, all neighbors should be in the same semantically similar space. Figure 5a shows an illustration of the $x_p$’s neighborhood of one image. We denote the semantically similar neighbor $x_q$ by circles and dissimilar neighbor $x_r$ by squares. If $x_r$ lies outside a smaller radius by a margin, as shown in Fig. 5b, we can reconstruct $x_p$ using all of its neighbors within that margin.

To obtain the semantically neighbors, the multiple-tag information should be incorporated into the distance metrics. Let us define $a$ and $b$ as two $m_i$-dimensional image feature vectors. We can rewrite measures such as L1, L2, and $\chi^2$ as a dot product of two vectors as follows:

$$\tilde{d}(a, b) = \sum_{i=1}^{n} w(i) \sum_{j=1}^{m_i} u'(j) \cdot d_{ij}(j)$$

where $\sum m_i = m, u'$ and $w$ are the weights. As shown in Fig. 5b, given the image $x_p$, along with its tag vector $y_p$, its target neighbor $x_q$ should be drawn closer, and $x_r$ should be pushed further away. Furthermore, for an image, the amount of push applied to its neighbors should vary depending on their conceptual similarity. Therefore, this metric learning problem can be formulated as follows:

$$\min_{w,u} \left\{ \sum_{pq} \eta_{pq} d(x_p, x_q) + \mu \sum_{pq} \eta_{pq} (1 - \lambda_{pr}) \right\}$$

where $\eta_{pq} = 1$ if $x_q$ is a target neighbor of $x_p$ and 0 otherwise, and $\lambda_{pq}$ and $\lambda_{pr}$ scale the error loss depending on the overlap between tag lists, which is defined as follows: $\lambda_{pq} = ||y_p \cap y_q|| / ||y_q||, \lambda_{pr} = \ldots$
\[ \| y_p \circ y_r \|_1 \| y_r \|_1, \mu \] is the controlling parameters, and \([z]_+ = \max(0, z)\) is the hinge loss. We solve Formula (6) using stochastic subgradient descent and projection steps (similar to Pegasos\(^{[39]}\)) to obtain an approximate optimal solution of \(w\) and \(u\).

For an image \(x_i\), we use the metric learned above to retrieve the \(k\) nearest neighbors. Together, these neighbors make up the local dictionary of image \(x_i\), which we denote as \(D_i = [x_{i1}, x_{i2}, \cdots, x_{ik}] \in \mathbb{R}^{m \times k}\).

We can formalize the reconstruction relationship as \(x_i = D_i \alpha_i + \zeta\), where \(\alpha_i \in \mathbb{R}^k\) is the coefficients vector for \(x_i\), \(\alpha_i(p) \geq 0\) and \(\sum_{p=1}^k \alpha_i(p) = 1\). Thus, we can obtain \(\alpha_i\) by solving the following problem:

\[
\min_{\alpha_i} \|\alpha_i\|_1 + \frac{\lambda}{2} \|x_i - D_i \alpha_i\|_2^2, \quad \text{s.t. } \alpha_i \geq 0 \tag{7}
\]

where \(\lambda\) is the regularization parameter that penalizes the disagreement between the reconstructed error and sparsity, \(\alpha_i = [\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{ik}]\), \(x_i = [x_{i1}, x_{i2}, \cdots, x_{ik}]^T\), \(\zeta^+\) is noise term, \(\zeta = \zeta^+ - \zeta^-\), \(\|\cdot\|_1\) is the \(L_1\) norm, \(D_i = \begin{bmatrix} D_{i1} & I_m & -I_m \\ E_{1 \times k} & 0_{1 \times m} & 0_{1 \times m} \end{bmatrix}\). We can efficiently solve Formula (7) using L1 optimization toolbox. Thus, we can represent \(x_i\) by a linear combination of neighbors and automatically determine the similarity between neighbors using the obtained reconstruction coefficient. Together, these neighbors make up the “semantically consistent neighborhood”.

As examples, we randomly selected semantically consistent neighbor images, which are shown in Fig. 6, and for comparison, we also show neighbors obtained by JEC\(^{[13]}\). From the figure, we can see that the neighbors returned from the “semantically consistent neighborhood” are not only visually alike but also partially visually correlated and semantically related.

### 3.3 Tag inference in semantic consistent neighborhood

We apply a semi-supervised image tagging algorithm to select a small number of the tags that are most probably semantically related to the image. We denote the predicted tag indicator matrix for all images as \(A\), which is split into two blocks as \(A = [A_L, A_U]^T\), where \(A_L\) is the tag indicator matrix containing predicted tag vectors of the first \(l\) training images and \(A_U\) contains the untagged vectors. We assume that the tag vector of one image can be reconstructed from the tags of its semantically consistent neighbors, and the coefficients are equal to the similarities of semantically consistent neighbors. We denote \(C = \{c_{ij}\}\) as the neighbor set similarity matrix, which is obtained using Formula (7). We can infer the tags of the untagged images by minimizing the reconstruction error as follows:

\[
\min_A \sum_{i=1}^n \|A_i - \sum_{j \neq i} c_{ij}A_j\|^2, \quad \text{s.t. } A_i = y_i, \quad \text{if } (x_i, y_i) \in L \tag{8}
\]

where \(y_i\) is the replenished tag vector of \(x_i\) obtained by Eq. (4). We can represent Formula (8) in matrix form as follows:

\[
\min_A \| (I - C) A^T (I - C) A \|_F \quad \text{s.t. } A = Y \tag{9}
\]

where \(Y = [y_1, y_2, \cdots, y_n]^T\) is the indicator matrix for replenished tags. If we set a derivative of the above equation with respect to \(A\) and set the obtained equation to zero, then we have the following:

\[
(H + H^T)A = QA = 0 \tag{10}
\]

where \(Q = H + H^T\) and \(H = (I - C)^T(I - C)\). By splitting \(Q\) into blocks after the \(l\)-th column and the \(l\)-th row, we have the following:

\[
Q = \begin{bmatrix} Q_{UL} & Q_{LU} \\
Q_{UL} & Q_{UU} \end{bmatrix} \tag{11}
\]

and we can then rewrite Eq. (10) as follows:

\[
Q_{UL}A_L + Q_{LU}A_U = 0, \quad Q_{UL}A_L + Q_{UU}A_U = 0 \tag{12}
\]

Thus, we can obtain the tag indicator matrix for untagged images by solving Eq. (13):

\[
Q_{UL}A_U = -Q_{UL}Y \tag{13}
\]

We can solve above equations by the Generalized Minimum Residual Method (GMRES)\(^{[40]}\).
4 Experiments

4.1 Datasets description

Four datasets, i.e., COREL5K\cite{13}, IAPR-TC12\cite{13}, ESPGAME\cite{13}, and MIRFLICKR-25000\cite{41} that have been used in previous works are used to evaluate the effectiveness of the proposed methods. For COREL5K dataset, we randomly selected 4500 images for training, and the rest for testing. For IAPR-TC12 dataset, we extracted nouns using part-of-speech tagger and discarded infrequent tags and grayscale images, then we selected 17,665 images for training. For ESPGAME dataset, we used 18,689 images for training. For MIRFLICKR-25000 dataset, we used half of images for training and half for testing.

Table 2 summarizes the statistical data of these datasets. We can see that around seventy percent of the tags have frequencies of less than the mean value. This is because these data are characterized by severe tag sparsity. We consider the large gap between the mean (or median) and maximum values of the “tags per image” to indicate the fact that many tags are missing.

In all datasets, each image is represented using color histogram features, gist features, and dense speedup robust features. More specifically, we extracted 44-dimensional color correlogram, 6-dimensional color moments, and 14-dimensional color texture moments, and then we concatenated them into a global feature. We used the L2 norm as the metric for gist feature, L1 for the global color histogram, and $\chi^2$ for local features.

4.2 Parameter tuning

In our method, we set four parameters in advance, which are tuned by a five-fold cross-validation, as shown in Figs. 7–10.

Specifically, we use parameter $\lambda$ to balance the weighted and smooth error in Eq. (1). We choose it from the pool $\{0.01, 0.1, 1, 10, 100\}$. From Fig. 7, we can see that the performance is stable when $\lambda$ ranges from 1 to 10. When $\lambda$ is too large, the smoothing error has greater weight, resulting in performance degradation. Similarly, too small a value of $\lambda$ results in overfitting of the training set. We set $\tau = 0.2$, $\delta = 20$, and $\mu = 1.0$. Taking the computational cost into account, we retrieved 100 samples with Formula (7).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Tags per image</th>
<th>Infrequent tags and ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>COREL5K</td>
<td>Mean 3.4</td>
<td>Median 4</td>
</tr>
<tr>
<td>IAPR-TC12</td>
<td>Mean 5.7</td>
<td>Median 5</td>
</tr>
<tr>
<td>ESPGAME</td>
<td>Mean 4.7</td>
<td>Median 5</td>
</tr>
<tr>
<td>MIRFLICKR-25000</td>
<td>Mean 8.9</td>
<td>Median 13</td>
</tr>
</tbody>
</table>
4.3 Performance evaluation with missing tags

First, we evaluated the capability of our method in dealing with the problem of tag sparsity. In this evaluation, we removed some tags randomly provided by users in a certain proportion and defined a missing rate, that equals the proportion of tags removed by users. For example, when a specific image had five original tags and the missing rate was 20%, we used only one tag for learning and then ranged the missing rate from 0.1 to 0.9. We compared the proposed method (SNL) with the following algorithms: SML [10], JEC [13], Tagprop [15], GS [16], and FastTag [19]. Figure 11 presents experimental results with COREL5K dataset. It is observed that there is performance degradation for all methods when missing rates varied from 0.1 to 0.9, and our proposed method has consistently better performance than others, which demonstrates the capability of SNL in dealing with missing tags.

To evaluate the SNL performance in each step, we tested different variants of SNL, the results of which are shown in Fig. 12. We can see that all of the steps are helpful. In detail, when predicting tags in a "semantically consistent neighborhood" (abbreviated as SNL(cn)), SNL achieved a marked improvement of 17% in F1 than that without tag replenishment (abbreviated as SNL(cn-r)). In addition, SNL(cn-r) outperformed SNL without "semantic balance neighborhood" (abbreviated as SNL(cn-r-b)) and SNL without metric learning (abbreviated as SNL(cn-r-b-m)).

4.4 Performance evaluation with noisy tags

To verify the capability of SNL in dealing with incorrect tags, we removed parts of original tags, and then randomly replenished the tags in the same proportion, ranging from 0.1 to 0.4, which we defined as a noisy rate. From the results shown in Fig. 13, we can see that the performances of all methods degraded with increases in the noisy rate. This is because noise leads directly to performance degradation. However, the performance of SNL outperforms all other methods because other methods depend heavily on the quality of associated tags. We can also see that TagProp performed better than FastTag and JEC. It is reasonable since TagProp predicts tags by per-tag model. Our proposed method is as straightforward as nearest neighbor voting. In this sense, FastTag and JEC are the methods most comparable to ours. Our method demonstrates a clear performance gain over them because they retrieve neighbor images only using visual features. These results clearly show the capability of SNL in dealing with noisy tags. To evaluate the performance of each step used in SNL, we tested different SNL variants, the results of which are shown in Fig. 14. In detail, we can see that SNL(cn) obtained a marked improvement of thirty percent in F1 against SNL(cn-r), SNL(cn-r-b), and SNL(cn-r-b-m). We can also see that all of the steps are helpful.
4.5 Performance evaluation with benchmark datasets

In Table 3, we show the performance of all methods on the different datasets measured by average precision (P), average recall (R), F1 value, and N+, thereby making it a fair comparison with the majority of the studies in this field of inquiry. First, we can see that TagProp and FastTag perform better than GS and SML. This can be explained by the fact that these methods are in fully supervised settings and their performance relies heavily on the tag quality. Second, we can see that GS performs better than JEC because sparsity prior and group clustering prior benefits the performance and stability of the model. Third, SNL method outperforms all these methods. This is because SNL applies tag inference in semantically consistent neighbors by exploring the diverse visual content and tags in the dataset.

Figure 15 shows a number of tagged images by the various methods. We can see that the proposed SNL method outperforms all the other methods.

5 Conclusion and Future Work

In this paper, we presented a novel semantic neighbor learning method from large-scale and noisy social datasets. Our key concept is the retrieval from big social data of neighbors that are not only visually alike and partially correlated but also semantically related, which facilitates the better transfer of tags between images and words, and which is a novel tag inference method for predicting tags that are robust to noise. To achieve this, we proposed a two-step method for learning the neighborhood from the dataset. Then, we proposed a semi-supervised tag inference method to assign to images a small number of the most probably semantically related tags. As a result, we can tag images more effectively using the socially tagged images as training data. To validate the effectiveness of our neighborhood learning method, we conducted extensive experiments on various datasets. Since it is effective and easy-to-use, we believe this proposed method to be of great practical use. In the future work, we plan to further improve the performance of the proposed method, by techniques that automatically set the value for the number of neighbors.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Nos. 61502094 and 61402099) and Natural Science Foundation of Heilongjiang Province of China (Nos. F2016002 and F2015020).

| Method | COREL5K P | COREL5K R | COREL5K F1 | COREL5K N+ | IAPR-TC12 P | IAPR-TC12 R | IAPR-TC12 F1 | IAPR-TC12 N+ | ESP-GAME P | ESP-GAME R | ESP-GAME F1 | ESP-GAME N+ | MIRFLICKR-25000 P | MIRFLICKR-25000 R | MIRFLICKR-25000 F1 | MIRFLICKR-25000 N+ |
|--------|-----------|-----------|-----------|-----------|-------------|-------------|-------------|-----------|-----------|-------------|-------------|----------------|---------------|----------------|----------------|
| SML    | 0.25      | 0.28      | 0.26      | 132       | 0.18        | 0.21        | 0.19        | 206       | 0.13      | 0.17        | 0.15        | 197           | 0.11       | 0.12        | 0.12        | 183 |
| JEC    | 0.26      | 0.32      | 0.29      | 137       | 0.28        | 0.29        | 0.29        | 223       | 0.19      | 0.21        | 0.20        | 219           | 0.15       | 0.16        | 0.156       | 272 |
| TagProp| 0.32      | 0.40      | 0.36      | 158       | 0.44        | 0.32        | 0.37        | 251       | 0.38      | 0.25        | 0.30        | 237           | 0.21       | 0.18        | 0.20        | 284 |
| GS     | 0.29      | 0.31      | 0.30      | 148       | 0.32        | 0.28        | 0.30        | 243       | 0.26      | 0.20        | 0.23        | 223           | 0.17       | 0.15        | 0.16        | 276 |
| FastTag| 0.31      | 0.34      | 0.324     | 152       | 0.45        | 0.24        | 0.31        | 279       | 0.47      | 0.21        | 0.30        | 246           | 0.22       | 0.17        | 0.19        | 293 |
| SNL    | 0.43      | 0.45      | 0.44      | 185       | 0.53        | 0.39        | 0.45        | 281       | 0.53      | 0.32        | 0.40        | 253           | 0.32       | 0.23        | 0.27        | 318 |
Fig. 15 Web image tagging examples (tags highlighted in bold are matching).

References


Xianmei Liu received the MS degree in computer science from Northeast Petroleum University, China in 2003. She is a full professor at the School of Computer and Information Technology, Northeast Petroleum University. She is an active research scientist in virtual reality and multimedia retrieval.

Maojun Cao received the MS degree in computer science from Northeast Petroleum University, China in 2006. He is currently an associate professor at the School of Computer and Information Technology in Northeast Petroleum University, China. His current research focuses on machine learning.
Feng Tian received the PhD degree in computer science from Beihang University in 2014, MS and BS degrees in computer science from Northeast Petroleum University, China in 2005 and 2002, respectively. He is currently an associate professor at the School of Computer and Information Technology in Northeast Petroleum University, China. He is currently a visiting scholar at the School Computing in National University of Singapore. His research interests include multimedia content understanding, information retrieval, and data mining. Various parts of his work have been published in ICME, Multimedia System, and Multimedia Tools and Applications. He has been served as reviewers for various journals and conferences.

Xukun Shen received the BS, MS, and PhD degrees in computer science from Beihang University, China, in 1987, 1994, and 2007, respectively. From 1998 to 2003, he was an associate professor with the College of Computer Science, Beihang University. He is currently a full professor with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His research interests include multimedia content management and computer graphics. Various parts of his work have been published in Computers & Graphics, The Visual Computer, and IEEE TIP. He has been served as reviewers for various journals and conferences.