# User-aware Image Tag Refinement via Ternary Semantic Analysis

Jitao Sang, Changsheng Xu, *Senior Member, IEEE*, Jing Liu, *Member, IEEE*

***Abstract*—Large-scale user contributed images with tags are easily available on photo sharing websites. However, the noisy or incomplete correspondence between the images and tags prohibits them from being leveraged for precise image retrieval and effective management. To tackle the problem of tag refinement, we propose a method of *Ranking based Multi-correlation Tensor Factorization* (RMTF), to jointly model the ternary relations among *user*, *image* and *tag*, and further to precisely reconstruct the user-aware image-tag associations as a result. Since the user interest or background can be explored to eliminate the ambiguity of image tags, the proposed RMTF is believed to be superior to the traditional solutions, which only focus on the binary image-tag relations. During the model estimation, we employ a ranking based optimization scheme to interpret the tagging data, in which the pair-wise qualitative difference between positive and negative examples is used, instead of the point-wise 0/1 confidence. Specifically, the positive examples are directly decided by the observed user-image-tag interrelations, while the negative ones are collected with respect to the most semantically and contextually irrelevant tags. Extensive experiments on a benchmark Flickr dataset demonstrate the effectiveness of the proposed solution for tag refinement. We also show attractive performances on two potential applications as the by-products of the ternary relation analysis.**

***Index Terms*—tag refinement, factor analysis, tensor factorization, social media**

## I. Introduction

With the popularity of Web 2.0 technologies, there are explosive photo sharing websites with large-scale image collections available online, such as Flickr,[1] Picasa,[2] Zooomr[3] and Pinterest.[4] These Web 2.0 websites allow users as owners, taggers, or commenters for their contributed images to interact and collaborate with each other in a social media dialogue. Its typical structure (Flickr as example) is illustrated in Fig.1, in which three types of interrelated entities are involved, i.e., *image*, *tag* and *user*. From this view, we can deem the user contributed tagging data as the products of the ternary interactions among images, tags and users.

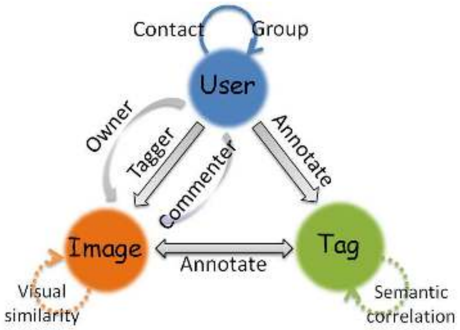


Fig. 1. An integrated structure of social tagging in Flickr

Obviously, given such a large-scale web dataset, noisy and missing tags are inevitable, which limits the performance of social tag-based retrieval system [1], [2]. Therefore, the tag refinement to denoise and enrich tags for images is desired to tackle this problem. Existing efforts on tag refinement [3], [4], [5], [6], [7], [8], [9], [10] exploited the semantic correlation between tags and visual similarity of images to address the noisy and missing issues, while the user interaction as one of important entities in the social tagging data is neglected.

As above mentioned, users are the originator of the tagging activity and they are involved with images and tags in many aspects. We believe that the incorporation of *user* information contributes to a better understanding and description of the tagging data. We take two simple examples to explain this observation. As shown in Fig.2(a), both images are tagged with “jaguar” by the two users (indicated by user ID,[5]) but they have different visual content, i.e., a luxury car and an animal respectively. Due to the well-known “semantic gap”, traditional work on image content understanding cannot solve the problem well. In this case, users’ interest and background information can be leveraged to specify the image semantics. That is, a car fan will possibly use “jaguar” to tag a ‘car’ image, while an animal specialist will use “jaguar” to tag a ‘wild cat’. Fig.2(b) shows three images from the FIFA 2010 final. We can see that different tags of “football” and “soccer” are annotated to the visually similar images. Considering the tagger information, we can easily understand this phenomenon: users have different tagging patterns. Maybe user *88077630@N00* is a Spanish fan while user *14915523@N05*

This work was supported in part by National Program on Key Basic Research Project (973 Program, Project No. 2012CB316304) and the National Natural Science Foundation of China (Grant No. 90920303, 61003161).

J. Sang, C. Xu (corresponding author) and J.Liu are with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; and also with the China-Singapore Institute of Digital Media, Singapore, 119613 (e-mail: jtsang@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn; jliu@nlpr.ia.ac.cn).

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

[2] http://picasa.google.com

[3] http://www.zooomr.com

[4] http://pinterest.com

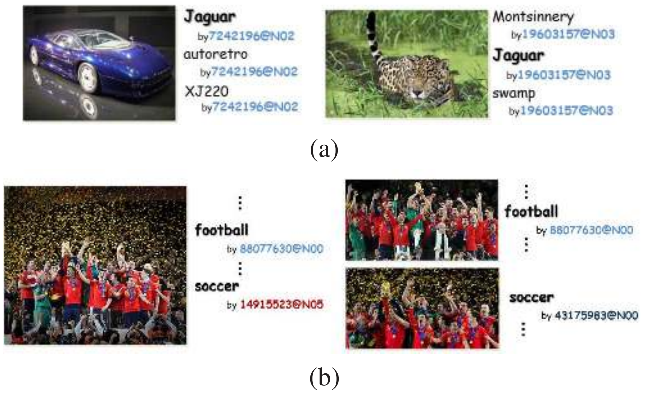
[5] The user ID of the taggers can be acquired from the Flickr API: http://www.flickr.com/services/api

and *43175983@N00* are Americans. These two examples can be considered as the reinforcement of tag understanding by introducing the *user* information. Note that it is not necessary to explicitly know the users' interests or profiles. What we are interested in is the fact that the tags are annotated by different users and there are variations in individual user's perspective and vocabulary. Incorporation of *user* may bring similar benefits to the image understanding. On top of visual appearance, the fact that images from the same user or tagged by similar users can capture more semantic correlations.

The goal of our work is to improve the underlying associations between the images and tags provided with the raw tagging data from photo sharing websites. To this end, in this paper, we solve it from a factor analysis perspective and aim at building the user-aware image and tag factor representations.[6] With the user factor incorporated, the image and tag factors will be free to focus on their own semantics and we can obtain more semantics-specified image and tag representations. A novel method named *Ranking based Multi-correlation Tensor Factorization* (RMTF) is proposed to tackle the tag refinement task. The framework is illustrated in Fig.3. It contains three primary parts: data collection, RMTF and tag refinement. For data collection, three types of data including users, images and tags as well as their ternary interrelations and intra-relations are collected.[7] In the RMTF module, we utilize tensor factorization to jointly model the multiple factors. To make full use of the observed tagging data and partial use of unobserved data, we present a novel ranking scheme for model estimation, which is based on the pair-wise qualitative difference between positive examples (i.e., observed tagging data) and negative ones (i.e., partial unobserved data). The collection of negative examples is carried out by analyzing user tagging behavior. The issue of noisy tags and missing tags are considered in a conservative filtering strategy by exploiting the tag correlation on context and semantics. Besides, the multiple intra-relations are employed as the smoothness constraints and then the factors inference is cast as a regularized tensor factorization problem. Finally, based on the learnt factor representations, which encode the compact users, images and tags representation over their latent subspaces, tag refinement is performed by computing the cross-space *image-tag* associations.

The main contributions of this paper are summarized as follows.

- We introduce *user* information into the social tag processing and jointly model the multiple factors of *user*, *image* and *tag* by 3-order tensor.
- We propose the RMTF model to extract the latent factor representations. A convergence provable learning algorithm is also presented.
- To make full use of the tagging data, a ranking optimization scheme is proposed to leverage the incomplete and ambiguous characteristics of user-generated tagging data.



(a)

(b)

Fig. 2. Example images from Flickr and their associated tags and taggers

- RMTF provides an entrance to other potential applications in the social media and information retrieval fields, which is discussed in *Section V.D*.

The rest of the paper is organized as follows. Related work is briefly reviewed in *Section* II. In *Section* III we formulate the problem and explain our basic idea. The detail of the proposed RMTF is addressed in *Section* IV. We report and discuss the experimental results as well as the applications in *Section* V. Finally, the conclusion and future work are given.

## II. Related Work

In this section, we briefly review some of the research literatures related to ternary analysis and image tag refinement.

### A. Ternary Analysis and Applications

Tensor is a mathematical representation of a multi-way array. The order of a tensor is the number of modes. A second-order tensor is a matrix, and a higher-order tensor has three or more modes. The most important tensor operation is tensor factorization. Many tensor factorization methods have been proposed, among which, CANDECOMP/PARAFAC (CP) and Tucker Decomposition are the most popular ones. A good survey for tensor factorization is provided in [11].

The advantage of ternary analysis is that we can use higher-order tensor to capture the multi-dimension relational data and employ tensor factorization to analyze their correlations. In the last decade, interest in ternary analysis has expanded to many fields, such as signal processing, numerical analysis, graphic analysis, and so on. We do not intend to cover all the related work and only focus on examples from the communities of computer vision and data mining.

In [12], the authors presented a dimensionality reduction algorithm based on tensor decomposition of N-mode SVD. They demonstrated the power of multilinear subspace analysis in the context of facial image ensembles. He *et al.* [13] also applied ternary analysis to the face recognition problem. Li *et al.* [14] introduced an online tensor subspace learning algorithm to the visual tracking problem. Considering the influence of the environment changing in the tracking process, Wen [15] extended the biased discriminant analysis (BDA) to

[6] These can be viewed as the feature matrices on the latent subspaces, which are spanned by the images and tags. We detail factor matrices derivation in *Section* 3

[7] We show a running example consisting of three users, five tags and four images in Fig.3(a).

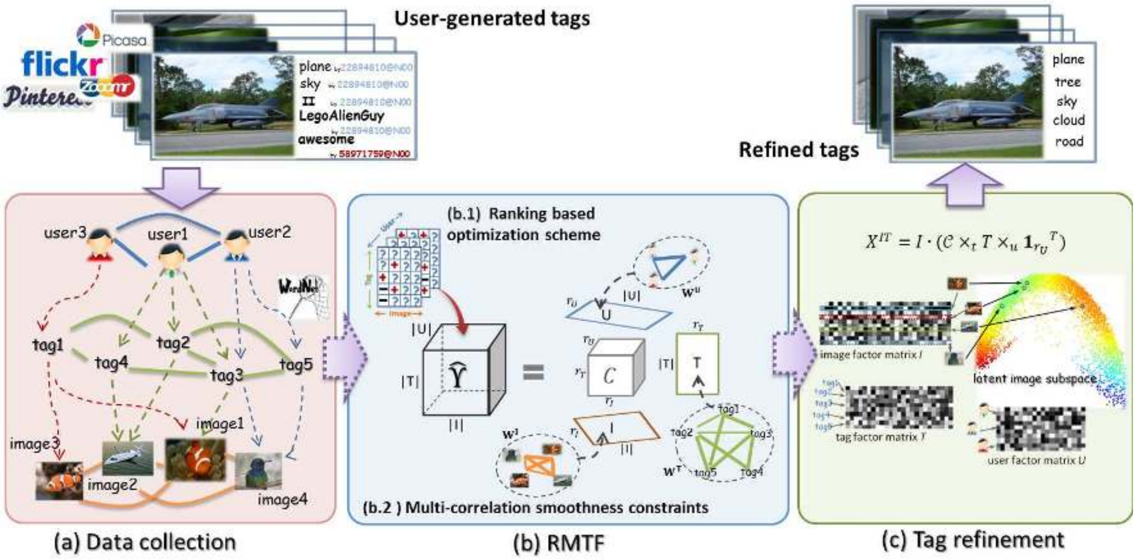


Fig. 3. The proposed framework.

the tensor biased discriminant analysis (TBDA) for appearance modeling and foreground extraction.

With the popularity of large-scale social media, massive amounts of data with multiple aspects and high dimensions are generated. Tensor provides a natural representation for such data. In [16], tensor factorization is utilized for multi-aspect data mining on the network data flow. Franz *et al.* [17] modeled the Semantic Web by a 3-dimensional tensor that enables the seamless representation of arbitrary semantic links. Under the rich media social network scenario, such as Diggs, Flickr, Last.fm, multi-relational data, user, item (post, photo), keyword, comment, contact are involved. In [18], the authors utilized 3-order tensor to model altogether users, tags and items in music sharing websites, and tackled personalized music recommendation based on latent semantic analysis. Beyond ternary relationship modeling, recently, researchers of [19] modeled the quaternary relationship among users, items, tags and ratings as a 4-order tensor and conducted multi-way latent semantic analysis. Lin *et al.* [20] introduced tensor factorization into photo sharing websites. The motivation is to extract meaningful communities by modeling the multi-relational social media contexts and interactions. As far as we know, little work has focused on incorporating user interaction to enhance the analysis of the correlation between images and tags. Our work is the first to incorporate user information into the task of image tag refinement.

One major challenge for ternary analysis is how to deal with the sparse and large-scale data. Standard tensor factorization methods do not account for the sparsity of the data. There have been substantial developments on variations of CP or Tucker Decomposition to account for the sparsity problem. For example, Kolda *et al.* [22] developed a greedy CP for sparse tensors that computes on triad at a time via an alternative least square (ALS) method. Recently, several works tackled the sparsity issue and avoided overfit by incorporating priors and combining with other schema. Chi *et al.* [23] employed the external information as smoothness priors into the tensor factorization and provided a good probabilistic interpretation. While in [24], the authors embeded a factorized representation of relations in a nonparametric Bayesian clustering framework, which achieves a tradeoff between the good predictive performance and interpretable representations. In this paper, the sparsity issue is addressed in two ways:

- The intra-relations among users, images and tags are employed as multi-correlation smoothness constraints into the tensor factorization model.
- We leverage the characteristics of user tagging activity and introduce a novel ranking optimization scheme.

### B. Image Tag Refinement

The literatures [10], [2] provide good surveys for the research work on image tag refinement. Along the structure of the tagging data illustrated in Fig.1, we characterize the related work according to the resources they leveraged.

As a pioneer work, Jin *et.al* [3] employed WordNet to estimate the semantic correlations among the annotated tags and remove weakly correlated ones. The work of [25] performed belief propagation among tags within the random walk with restart framework to refine the imprecise original annotations. In [6], Xu *et al.* proposed to jointly model the tag similarity and tag relevance and perform tag refinement from the topic modeling view. These work is typically based on the *tag-tag* analysis. In [26], the authors explicitly considered the tag-image and tag-tag relations and proposed a dual cross-media relevance model for image annotation. Liu *et al.* [5] proposed to rank the image tags according to their relevance w.r.t. the associated images by modeling tag similarity and image similarity. In [9], the improved tag assignments are learnt by maximizing the consistency between visual similarity and semantic similarity while minimizing the deviation from initially user-provided tags. An interesting work is done by Xie *et al.*[27], in which several important issues in building an end-to-end image tagging application are addressed, including tagging vocabulary design, taxonomy-based tag refinement, classifier score calibration for tag ranking, and selection of valuable tags. Recently, Liu *et al.*[28] proposed a multi-edge graph based unified framework to solve the image annotation, tag-to-region and tag refinement problem. *Tag-tag*, *image-image* and *image-tag* relationships are explored in these work.

The most related work to this paper is [7], [10], which solves the tag refinement problem through low-rank matrix approximation. Zhu *et al.* [10] considered the tagging characteristics from the view of low-rank, error sparsity, content consistency and tag correlation. In [7], a factor analysis model is proposed and the tag refinement problem is cast as estimating the image-tag correlations. While these work simultaneously modeled the *tag-tag*, *image-image* and *image-tag* relationships, they aggregated images' tags over all users, thereby losing important information about individual user's variation in tag usage. In this paper, we exploit the social aspect of the photo sharing websites and consider *user* factor into the tag refinement problem. We believe that incorporation of *user* information will facilitate explaining the tagging data and lead to better estimates of image and tag factors.

## III. Problem Formulation

The low dimensional *user*, *image* and *tag* factor matrices can be viewed as compact representations in the corresponding latent subspaces. The latent subspaces capture the relevant attributes, e.g., the user dimensions are related to users' preferences or social interests, the image dimensions indicate visual themes and the tag dimensions are related to the semantic topics of tags. The basic intuition behind this work is: *The incorporation of user information will help extract more compact and informative image and tag representations in the semantic subspaces. The task of image tag refinement is then solved by computing the cross-space image-tag associations.* In this section we first introduce the idea of jointly modeling the *user*, *image* and *tag* factors into a tensor factorization framework, then explain how to employ the derived factors for tag refinement.

In the following, we denote tensors by calligraphic uppercase letters (e.g., $\mathcal{Y}$), matrices by uppercase letters (e.g., $U, I, T$), vectors by bold lowercase letters (e.g., $\boldsymbol{u}, \boldsymbol{i}$), scalars by lowercase letters (e.g., $u, i$) and sets by blackboard bold letters (e.g., $\mathbb{U}, \mathbb{I}, \mathbb{T}$).

### A. Tensor Factorization

There are three types of entities in the photo sharing websites. The tagging data can be viewed as a set of triplets. Let $\mathbb{U}, \mathbb{I}, \mathbb{T}$ denote the sets of users, images, tags and the set of observed tagging data is denoted by $\mathbb{O} \subset \mathbb{U} \times \mathbb{I} \times \mathbb{T}$, i.e., each triplet $(u, i, t) \in \mathbb{O}$ means that user $u$ has annotated image $i$ with tag $t$. For example, the left image in Fig.2(a) corresponds to three triplets in $\mathbb{O}$ sharing the same image and user. The ternary interrelations can be viewed as a three-mode cube, where the modes are the *user*, *image* and *tag*. Therefore, we can induce a three dimensional tensor $\mathcal{Y} \in \mathbb{R}^{|\mathbb{U}|\times|\mathbb{I}|\times|\mathbb{T}|}$, which is defined as:

$$y_{u,i,t} = \begin{cases} 1 & \text{if } (u,i,t) \in \mathbb{O} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $|\mathbb{U}|$, $|\mathbb{I}|$, $|\mathbb{T}|$ are the number of distinct users, images and tags respectively. Fig.5(a) shows the tensor constructed from the running example base on Eq.1.

To jointly model the three factors of *user*, *image* and *tag*, we employ the general tensor factorization model, Tucker Decomposition for the latent factor inference. In Tucker Decomposition, the tagging data $\mathcal{Y}$ are estimated by three low rank matrices and one core tensor (see Fig.4):

$$\hat{\mathcal{Y}} := \mathcal{C} \times_u U \times_i I \times_t T \tag{2}$$

where $\times_n$ is the tensor product of multiplying a matrix on mode $n$. Each low rank matrix ($U \in \mathbb{R}^{|\mathbb{U}|\times r_U}$, $I \in \mathbb{R}^{|\mathbb{I}|\times r_I}$, $T \in \mathbb{R}^{|\mathbb{T}|\times r_T}$) corresponds to one factor. The core tensor $\mathcal{C} \in \mathbb{R}^{r_U \times r_I \times r_T}$ contains the interactions between the different factors. The ranks of decomposed factors are denoted by $r_U, r_I, r_T$ and Eq.2 is called *rank-*$(r_U, r_I, r_T)$ Tucker decomposition. An intuitive interpretation of Eq.2 is that the tagging data depend not only on how similar an image's visual features and tag's semantics are, but on how much these features/semantics match with the users' preferences.

Typically, the latent factors $U$, $I$, $T$ can be inferred by directly approximating $\mathcal{Y}$ and the tensor factorization problem is reduced to minimizing an point-wise loss on $\hat{\mathcal{Y}}$:

$$\min_{U,I,T,\mathcal{C}} \sum_{(\tilde{u},\tilde{i},\tilde{t})\in|\mathbb{U}|\times|\mathbb{I}|\times|\mathbb{T}|} (\hat{y}_{\tilde{u},\tilde{i},\tilde{t}} - y_{\tilde{u},\tilde{i},\tilde{t}})^2 \tag{3}$$

where $\hat{y}_{\tilde{u},\tilde{i},\tilde{t}} = \mathcal{C} \times_u \boldsymbol{u}_{\tilde{u}} \times_i \boldsymbol{i}_{\tilde{i}} \times_t \boldsymbol{t}_{\tilde{t}}$. As this optimization scheme tries to fit to the numerical values of 1 and 0, we refer it as the *0/1 scheme*. To alleviate the sparse problem and better utilize the tagging data, in this paper, we propose RMTF for factor inference, which is detailed in *section* IV.

### B. Tag Refinement

From the perspective of subspace learning, the derived factor matrices $U, I, T$ can be viewed as the feature representations on the latent *user*, *image*, *tag* subspaces, respectively. As illustrated in Fig.3(c), each row of the factor matrices corresponds to one object (user, image or tag). The core tensor $\mathcal{C}$ defines a multi-linear operation and captures the interactions among different subspaces. Therefore, multiplying a factor matrix to the core tensor is related to a change of basis. We define

$$\mathcal{T}^{UI} := \mathcal{C} \times_t T \tag{4}$$

then $\mathcal{T}^{UI} \in \mathbb{R}^{r_U \times r_I \times |\mathbb{T}|}$ can be explained as the tags' feature representations on the $user \times image$ subspace. Each $r_U \times r_I$ slice of matrix corresponds to one tag feature representation. By summing $\mathcal{T}^{UI}$ over the $user$ dimensions, we can obtain the tags' representations on the $image$ subspace. Therefore, the cross-space image-tag association matrix $X^{IT} \in \mathbb{R}^{|\mathbb{I}|\times|\mathbb{T}|}$ can be calculated as:[8]

$$X^{IT} = I \cdot (\mathcal{T}^{UI} \times_u \mathbf{1}_{r_U}^{\top}) \tag{5}$$

[8] In practice, for new images not in the training dataset, we can approximate their positions in the learnt image subspace by using approximated eigenfunctions based on the kernel trick [29].

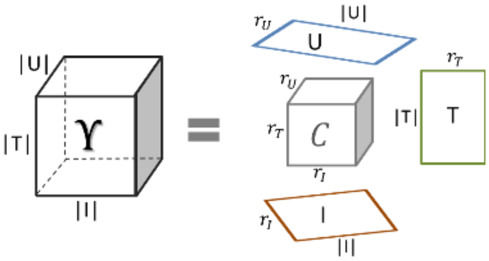


Fig. 4. Tucker decomposition: the tensor $\mathcal{Y}$ is constructed by multiplying three factor matrices $U, I, T$ to a small core tensor $\mathcal{C}$.

The tags with the $K$ highest associations to image $i$ are reserved as the final annotations:

$$Top(i, K) = \max_{t \in \mathbb{T}}{}^{K} X^{IT}_{i:} \tag{6}$$

In the experiment, we fix $K = 10$.

## IV. Ranking based Multi-correlation Tensor Factorization

In this section, we detail the proposed RMTF model. We first introduce a novel ranking based optimization scheme for better interpretation of the tagging data. Then the multiple intra-relations among users, images and tags are utilized as the smoothness constraints for the latent factors and finally we present a convergence provable learning algorithm.

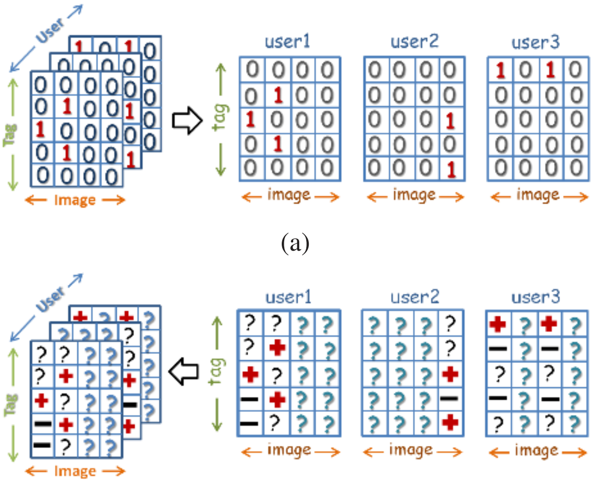
### *A. Ranking based Optimization Scheme*

Traditional factorization models [7], [10] approximate the tagging data based on the *0/1 scheme*. Under the situation of social image tagging data, the semantics of encoding all the unobserved data as 0 are incorrect, which is illustrated with the running example of Fig.3(a):

- Firstly, the fact that *user3* has not given any tag to *image2* and *image4* does not mean that *user3* considered all the tags are bad for describing the images.[9] Maybe he/she does not want to annotate the image or has no chance to see the image.
- Secondly, *user1* annotates *image1* with only *tag3*. It is also unreasonable to assume that other tags should not be annotated to the image, as some concepts may be missing in the user-generated tags and individual user may not be familiar to all the relevant tags in the large tag set.

According to the optimization function in Eq.3, the learning process tries to predict 0 for both cases, which is apparently unreasonable. To address the above problems, we present a ranking optimization scheme which intuitively considers the user tagging behaviors and addresses the issues of missing tags and noisy tags.

We note that only the qualitative difference is important and fitting to the numerical values of 1 and 0 is unnecessary. Therefore, instead of solving an point-wise classification task, we formulate it as a ranking problem which uses tag pairs within each user-image combination $(u, i)$ as the training data and optimizes for correct ranking. For example, $y(u, i, t^+) > y(u, i, t^-)$ indicates that user $u$ considers tag $t^+$ is better to describe image $i$ than tag $t^-$.



user1
0 0 0 0
0 1 0 0
1 0 0 0
0 1 0 0
0 0 0 0

user2
0 0 0 0
0 0 0 0
0 0 0 1
0 0 0 0
0 0 0 1

user3
1 0 1 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0

(a)

user1
? ? ? ?
? + ? ?
+ ? ? ?
− + ? ?
− ? ? ?

user2
? ? ? ?
? ? ? ?
? ? ? +
? ? ? −
? ? ? +

user3
+ ? + ?
− ? − ?
? ? ? ?
− ? − ?
? ? ? ?

(b)

Fig. 5. Tagging data interpretation. (a) 0/1 scheme (b) ranking scheme

We provide some notations for easy explanation. Each user-image combination $(u, i)$ is defined as a *post*. The set of observed posts is denoted as $\mathbb{P}_{\mathbb{O}}$:

$$\mathbb{P}_{\mathbb{O}} = \{(u, i) | \exists t \in \mathbb{T}, y_{u,i,t} = 1\} \tag{7}$$

The neutral triplets constitute a set $\mathbb{M}$:

$$\mathbb{M} = \{(u, i, t) | (u, i) \notin \mathbb{P}_{\mathbb{O}}\} \tag{8}$$

It is arbitrary to treat the neutral triplets as either positive or negative and we remove all the triplets in $\mathbb{M}$ from the learning process (filled by bold question marks in Fig.5(b)).

For the training pair determination, we consider two characteristics of the user tagging behaviors. On one hand, some concepts maybe missing in the user-generated tags. We assume that the tags co-occurring frequently are likely to appear in the same image (we call it *context-relevant*). On the other hand, users will not bother to use all the relevant tags to describe the image. The tags *semantic-relevant* with the observed tags are also the potential good descriptions for the image. The two assumptions are reasonable. Looking at the running example, *user1* annotated *image1* with *tag3* (we assume *tag3* is to describe Nemo, e.g., *tag3*="fish"). We can see that the tags "water", "sea", "coral" which are *context-relevant* and "animal", "seafish" "clownfish" which are *semantic-relevant* with the tag "fish" are all good descriptions for *image1*. To perform the idea, we build a tag affinity graph $W^T$ based on tag semantic and context intra-relations.[10] The tags with the $k$-highest affinity values are considered semantic-relevant or context-relevant.

Regarding the possible noises in the user-generated tags, it is risky to enrich the semantic or context relevant tags into the positive set. Therefore, we choose a conservative strategy: we keep the unobserved tags semantic-**irrelevant** and context-**irrelevant** with any of the observed tags, to form the negative

[9] We call triplets like $(u_3, i_2, :)$ and $(u_3, i_4, :)$ as the neutral triplets.

[10] Detail of $W^T$ construction is introduced in next subsection.

tag set. Note that the ranking optimization is performed over each post and within each post $(u, i)$ a positive tag set $\mathbb{T}^+_{u,i}$ and a negative tag set $\mathbb{T}^-_{u,i}$ are desired to construct the training pairs. Given a post $(u, i) \in \mathbb{P}_{\mathbb{O}}$, the observed tags constitute a positive tag set (the corresponding triplets are filled by plus signs in Fig.5(b)):

$$\mathbb{T}^+_{u,i} = \{t|(u, i) \in \mathbb{P}_{\mathbb{O}} \wedge y_{u,i,t} = 1\} \tag{9}$$

The negative tag set is constituted as:

$$\mathbb{T}^-_{u,i} = \left\{t|(u, i) \in \mathbb{P}_{\mathbb{O}} \wedge y_{u,i,t} \neq 1 \wedge t \notin \mathbb{N}_{\mathbb{T}^+_{u,i}}\right\} \tag{10}$$

where $\mathbb{N}_{\mathbb{T}^+_{u,i}}$ indicates the set of tags relevant to the annotated tags in post $(u, i)$. Then $t_4, t_5 \in \mathbb{T}^-_{u_1,i_1}$, presumably $tag1$ and $tag2$ are relevant to $tag3$. The final tagging data representation for the running example is illustrated in Fig.5(b). The triplets corresponding to tags $t \in \mathbb{N}_{\mathbb{T}^+_{u,i}}$ are also removed from the learning process and filled by plain question marks. The minus signs indicate the filtered negative triplets.

Any tag $t \in \mathbb{T}^+_{u,i}$ is considered a better description for image $i$ than all the tags $t \in \mathbb{T}^-_{u,i}$. The pairwise ranking relationships can be denoted as:

$$\hat{y}_{u,i,t_1} > \hat{y}_{u,i,t_2} \Leftrightarrow t_1 \in \mathbb{T}^+_{u,i} \wedge t_2 \in \mathbb{T}^-_{u,i} \tag{11}$$

The optimization criterion is to minimize the violation of the pairwise ranking relationships in the reconstructed tensor $\hat{\mathcal{Y}}$, which leads to the following objective:

$$\min_{U,I,T,\mathcal{C}} \sum_{(\tilde{u},\tilde{i})\in\mathbb{P}_{\mathbb{O}}} \Big( \sum_{t^+\in\mathbb{T}^+_{\tilde{u},\tilde{i}}} \sum_{t^-\in\mathbb{T}^-_{\tilde{u},\tilde{i}}} f(\hat{y}_{\tilde{u},\tilde{i},t^-} - \hat{y}_{\tilde{u},\tilde{i},t^+})\Big) \tag{12}$$

where $f : \mathbb{R} \to [0, 1]$ is a monotonic increasing function (e.g., the logistic sigmoid function or Heaviside function). Through necessary algebra manipulation, we derive the matrix form of the objective function:

$$\min_{U,I,T,\mathcal{C}} f \left( \begin{array}{c} \vdots \\ \mathcal{C} \times_u \boldsymbol{u}_{\tilde{u}} \times_i \boldsymbol{i}_{\tilde{i}} \times_t (T^-_{\tilde{u},\tilde{i}} \otimes \mathbf{1}^\top_{|\mathbb{T}^-_{\tilde{u},\tilde{i}}|} - T^+_{\tilde{u},\tilde{i}} \otimes \mathbf{1}^\top_{|\mathbb{T}^+_{\tilde{u},\tilde{i}}|}) \\ \vdots \end{array} \right) \times \mathbf{1}_{\sum_{(\tilde{u},i)\in\tilde{\mathbb{P}}_{\mathbb{O}}|\mathbb{T}^+_{\tilde{u},\tilde{i}}|\cdot|\mathbb{T}^-_{\tilde{u},\tilde{i}}|}}$$

where $\otimes$ is the cross product, $f$ switches to a component-wise function and $\mathbf{1}_D \in \mathbb{R}^{1\times D}$ is 1-vector with all the elements $\mathbf{1}_d = 1$. $\mathbb{T}^+_{\tilde{u},\tilde{i}}$ is the positive tag set for the post $(\tilde{u}, \tilde{i})$:

$$\mathbb{T}^+_{\tilde{u},\tilde{i}} = \left\{t_1^{(\tilde{u},\tilde{i})^+}, \cdots, t_{|\mathbb{T}^+_{\tilde{u},\tilde{i}}|}^{(\tilde{u},\tilde{i})^+}\right\}$$

$T^+_{\tilde{u},\tilde{i}} \in R^{|\mathbb{T}^+_{\tilde{u},\tilde{i}}|\times r_T}$ is the tag vector matrix composed by the positive tags in $\mathbb{T}^+_{\tilde{u},\tilde{i}}$: $T^+_{\tilde{u},\tilde{i}} = \left(\boldsymbol{t}^\top_{(\tilde{u},\tilde{i})^+:1}, \cdots, \boldsymbol{t}^\top_{(\tilde{u},\tilde{i})^+:|\mathbb{T}^+_{\tilde{u},\tilde{i}}|}\right)^\top$. Here $\boldsymbol{t}_{(\tilde{u},\tilde{i})^+:\tilde{t}}$ is $t_{\tilde{t}}^{(\tilde{u},\tilde{i})^+}$-th row vector of the tag factor matrix.

Note that the number of positive and negative tags in the post $(\tilde{u}, \tilde{i})$, $|\mathbb{T}^+_{\tilde{u},\tilde{i}}|$ and $|\mathbb{T}^-_{\tilde{u},\tilde{i}}|$, are constant once the tag relevances are determined. For simplicity, we denote $N = \sum_{(\tilde{u},\tilde{i})\in\mathbb{P}_{\mathbb{O}}} |\mathbb{T}^+_{\tilde{u},\tilde{i}}| \cdot |\mathbb{T}^-_{\tilde{u},\tilde{i}}|$ and further define

$$\mathbf{p}^\top = \left( \begin{array}{c} \vdots \\ \mathcal{C} \times_u \boldsymbol{u}_{\tilde{u}} \times_i \boldsymbol{i}_{\tilde{i}} \times_t (T^-_{\tilde{u},\tilde{i}} \otimes \mathbf{1}^\top_{|\mathbb{T}^-_{\tilde{u},\tilde{i}}|} - T^+_{\tilde{u},\tilde{i}} \otimes \mathbf{1}^\top_{|\mathbb{T}^+_{\tilde{u},\tilde{i}}|}) \\ \vdots \end{array} \right)$$

$\mathbf{p}$ is a long row vector of length $\sum_{(\tilde{u},\tilde{i})\in\mathbb{P}_{\mathbb{O}}} |\mathbb{T}^+_{\tilde{u},\tilde{i}}| \cdot |\mathbb{T}^-_{\tilde{u},\tilde{i}}|$. Therefore, with our novel ranking optimization scheme, the tucker decomposition model amounts to minimizing:

$$f(\mathbf{p}^\top) \times \mathbf{1}_N \tag{13}$$

Note that the work in [30], [31] provided similar ranking schemes for recommender systems, while the main difference is that we explicitly consider the incomplete and ambiguous characteristics of the user-generated tagging data and filter out the quasi-positive tags. In their formulation, given a post $(u, i) \in \mathbb{P}_{\mathbb{O}}$, all the tags that not annotated by *user u* to *image i* will be treated as negative tags, and the corresponding negative set is:

$$\mathbb{T}^-_{u,i} = \{t|(u, i) \in \mathbb{P}_{\mathbb{O}} \wedge y_{u,i,t} \neq 1\} \tag{14}$$

Apparently, this formulation ignores the issues of missing tags and noisy tags, which cannot be directly applied to the social tagging problems. In addition, Rendle employed l-1 norm for regularization, while in the proposed RMTF, additional multiple intra-relations are utilized as the smoothness constraints, which is detailed in the following subsection.

### B. Multi-correlation Smoothness Constraints

In addition to the ternary interrelations, we also collect multiple intra-relations among users, images and tags. These intra-relations constitute the user, image, tag affinity graphs $W^U \in \mathbb{R}^{|\mathbb{U}|\times|\mathbb{U}|}$, $W^I \in \mathbb{R}^{|\mathbb{I}|\times|\mathbb{I}|}$ and $W^T \in \mathbb{R}^{|\mathbb{T}|\times|\mathbb{T}|}$, respectively. Two objects with high affinities should be mapped close to each other in the learnt subspaces. Therefore, the intra-relations are employed as the smoothness constraints to preserve the affinity structure in the low dimensional factor subspaces. In this subsection, we first introduce how to construct the affinity graphs, and then incorporate them into the tensor factorization framework.

**User affinity graph $W^U$.** Generally speaking, the activity of joining in interesting groups indicates the users' interests and backgrounds. Also, the group statistic is more easy to obtain compared with other privacy concerning information, e.g., searching history, the query log, etc. Therefore, we measure the affinity relationship between user $u_m$ and $u_n$ using the co-occurrence of their joined groups:

$$W^U_{m,n} = \frac{n(u_m, u_n)}{n(u_m) + n(u_n)} \tag{15}$$

where $n(u_m)$ is the number of groups user $u_m$ joined and $n(u_m, u_n)$ is the number of groups $u_m$ and $u_n$ co-joined.

**Image affinity graph $W^I$.** To measure the visual similarities between images, each image is extracted a 428-dimensional feature vector $\boldsymbol{d}$ as the visual representation [10], [9], including 225-d blockwise color moment features, 128-d wavelet texture features and 75-d edge distribution histogram features. The image affinity graph $W^I$ is defined based on the following Gaussian RBF kernel:

$$W^I_{m,n} = e^{-||\boldsymbol{d}_m - \boldsymbol{d}_n||^2/\sigma_I^2} \tag{16}$$

where $\sigma_I$ is set as the median value of the elements in $W^I$.

**Tag affinity graph $W^T$.** To serve the ranking based optimization scheme, we build the tag affinity graph based on the tag context and semantic relevance. The context relevance of tag $t_m$ and $t_n$ is simply encoded by their weighted co-occurrence in the image collection:

$$t^c_{m,n} = \frac{n(t_m, t_n)}{n(t_m) + n(t_n)} \tag{17}$$

For tag semantic relevance, we follow Liu *et al.* [9]'s approach and estimate the semantic relevance between tag $t_m$ and $t_n$ based on their WordNet distance:

$$t^s_{m,n} = \frac{2 \cdot IC(lcs(t_m, t_n))}{IC(t_m) + IC(t_n)} \tag{18}$$

where $IC(\cdot)$ is the information content of tag, and $lcs(t_i, t_j)$ is their least common subsumer in the WordNet taxonomy. The tag affinity graph is constructed as:

$$W^T_{m,n} = \lambda_c t^c_{m,n} + \lambda_s t^s_{m,n} \tag{19}$$

where $\lambda_c + \lambda_s = 1$, $\lambda_c$ and $\lambda_s$ are the weights of context relevance and semantic relevance.[11] Note that we have no requirements on how to build the affinity graphs and other intra-relation measurements can also be explored.

The affinity graphs are utilized as the regularization terms to impose smoothness constraints for the latent factors. All the affinity graphs are normalized. Take the image affinity graph $W^I$ as an example, the regularization term is:

$$\sum_{m=1}^{|\mathbb{I}|} \sum_{n=1}^{|\mathbb{I}|} W^I_{m,n} ||\boldsymbol{i_m} - \boldsymbol{i_n}||^2 \tag{20}$$

where $||\cdot||^2$ denotes the Frobenius norm. The basic idea is to make the latent representations of two images as close as possible if there exists strong affinity between them. We can achieve this by minimizing $tr(I^\top L_I I)$, where $tr(\cdot)$ denotes the trace of a matrix and $L_I$ is the Laplacian matrix for the image affinity matrix $W^I$. Similar regularization terms can be added for the user and tag factors. In this way, the extracted data characteristics are consistent with such prior knowledge, which alleviate the sparsity problem as well as control over the outcomes.

Combining with Eq.13, we obtain the overall objective function:

$$\begin{aligned} \min_{U,I,T,\mathcal{C}} g = f(\mathbf{p}^\top) \times \mathbf{1}_N + \beta(||\mathbb{U}||^2 + ||\mathbb{I}||^2 + ||\mathbb{T}||^2) \\ + \alpha(tr(U^\top L_U U) + tr(I^\top L_I I) + tr(T^\top L_T T)) \end{aligned} \tag{21}$$

where $||\mathbb{U}||^2 + ||\mathbb{I}||^2 + ||\mathbb{T}||^2$ is *l*-1 regularization term to penalize large parameters, $\alpha$ and $\beta$ are weights controlling the strength of corresponding constraints.

**Algorithm 1** Alternating Learning Algorithm

```
Input:
    User tagging tensor $\mathcal{Y} \in \mathbb{R}^{|\mathbb{U}|\times|\mathbb{I}|\times|\mathbb{T}|}$; affinity graph adjacency
    matrices $W^U \in \mathbb{R}^{|\mathbb{U}|\times|\mathbb{U}|}$, $W^I \in \mathbb{R}^{|\mathbb{I}|\times|\mathbb{I}|}$, $W^T \in \mathbb{R}^{|\mathbb{T}|\times|\mathbb{T}|}$;
    rank of the factor matrices $r_U$, $r_I$, $r_T$; and the weighting
    parameters $\alpha$, $\beta$.
Output:
    User, image and tag factor matrices $U \in \mathbb{R}^{|\mathbb{U}|\times r_U}$, $I \in \mathbb{R}^{|\mathbb{I}|\times r_I}$,
    $T \in \mathbb{R}^{|\mathbb{T}|\times r_T}$ and the core tensor $\mathcal{C} \in \mathbb{R}^{r_U\times r_I\times r_T}$.
initialize random dense matrices $U^{(0)} \in \mathbb{R}^{|\mathbb{U}|\times r_U}$, $I^{(0)} \in$
    $\mathbb{R}^{|\mathbb{I}|\times r_I}$, $T^{(0)} \in \mathbb{R}^{|\mathbb{T}|\times r_T}$; $t \leftarrow 0$.
repeat
    $\mathcal{C}^{(t+1)} = \arg\min g(U^{(t)}, I^{(t)}, T^{(t)}, \mathcal{C})$
    $\mathcal{U}^{(t+1)} = \arg\min g(U, I^{(t)}, T^{(t)}, \mathcal{C}^{(t+1)})$
    $\mathcal{I}^{(t+1)} = \arg\min g(U^{(t+1)}, I, T^{(t)}, \mathcal{C}^{(t+1)})$
    $\mathcal{T}^{(t+1)} = \arg\min g(U^{(t+1)}, I^{(t+1)}, T, \mathcal{C}^{(t+1)})$
    $t \leftarrow t + 1$
until converge
return $U = U^{(t-1)}$,$I = I^{(t-1)}$,$T = T^{(t-1)}$,$\mathcal{C} = \mathcal{C}^{(t-1)}$
```

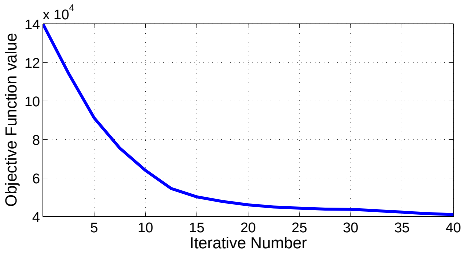


Fig. 6. The convergence curve of Algorithm 1.

### *C. Learning Algorithm*

Next we present an algorithm to solve the optimization problem. Obviously, directly optimizing Eq.21 is infeasible and we use an iterative optimization algorithm. To begin with, we first provide the following theorem:

**Theorem 1**. *$g$ is strictly convex w.r.t. $U$, $I$, $T$ and $\mathcal{C}$, respectively.*

We propose an alternating learning algorithm (ALA) to learn the factors by iteratively optimizing each subproblems, which is shown in Algorithm 1. According to Theorem 1, each subproblem in Algorithm 1 has a unique solution. In practise, as $g$ is convex w.r.t. $I$, it is also convex w.r.t. each $\boldsymbol{i}_m$.[12] Therefore, when performing optimization on $I$, we optimize one row $\boldsymbol{i}_m$ at a time with other rows $\{\boldsymbol{i}_1, \cdots, \boldsymbol{i}_{m-1}, \boldsymbol{i}_{m+1}, \cdots, \boldsymbol{i}_{r_I}\}$ fixed. We prove that the learning algorithm has a good convergence property.

**Theorem 2**. *The alternating learning algorithm converges to a local optimum.*

The proof of Theorem 1 directly follows the regularized matrix factorization [32] and is omitted here. We provide the proof of Theorem 2 in *Appendix* A. With the learnt factors, tag refinement is performed by computing the cross-space *image-tag* associations as discussed in *Section* III.B.

In the experiments, we observed that the proposed ALA converges to the minimum after about 20 iterations. Fig.6

[11] In the experiment, we choose $\lambda_c = 0.9$ and $\lambda_s = 0.1$.

[12] The user factor $U$ and tag factor$T$ are the same cases as the image factor $I$.

TABLE I
THE STATISTICS OF NUS-WIDE-USER15

| | Users $\|\mathbb{U}\|$ | Images $\|\mathbb{I}\|$ | Tags $\|\mathbb{T}\|$ | $\|\mathbb{O}\|$ |
|---|---|---|---|---|
| USER15 | 3,372 | 124,099 | 5,018 | 1,223,254 |

shows the change of objective function values in the convergence process. We perform our experiments on MATLAB in a PC with 2.13GHz CPU and 16 GB memory. The convergence time on the experimental dataset is about 6 hours. Actually, in the proposed learning algorithm, each factor vector $\boldsymbol{i}_m$ is updated independently of other vectors, which gives rise to potentially massive parallelization (e.g. parallel MATLAB). Theoretically, the algorithm achieves a linear converge speedup which is proportion to the number of used processors [33]. *Distributed* storing also provides a convenient way to store very large matrices. The larger $r_U$, $r_I$, and $r_T$ are, the more obviously the speedup is.

Note that the user, image and tag factor matrices are initialized randomly in the proposed learning algorithm. Likewise to other non-convex learning problems, the initialization of the factor matrices is very important to our learning algorithm. We will be working towards investigating a proper initialization scheme in the future.

## V. Experiments

### A. Data Set

We perform the experiments of social tag refinement on the large-scale web image dataset, NUS-WIDE [1]. It contains 269,648 images with 5,018 unique tags collected from Flickr. We crawled the owner information according to the image ID and obtained the owner user ID of 247,849 images.[13] The collected images belong to 50,120 unique users, with each user owning about 5 images. We select the users owning no less than 15 images and keep their images to obtain our experimental dataset, which is referred as NUS-WIDE-USER15. Table I summarizes the collected dataset. $|\mathbb{O}|$ is the number of observed triplets. The NUS-WIDE provides ground-truth for 81 tags of the images. In the experiments, we evaluate the performance of tag refinement by the F-score metric:

$$Fscore = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{22}$$

### B. Impact of Parameters

The proposed approach, RMTF, has five parameters, the rank of factor matrices $r_U$, $r_I$, $r_T$ and the regularization weights $\alpha$, $\beta$. We explore the influence of different parameter settings on a smaller but representative dataset, NUS-WIDE-USER50, which has 588 users and 55,141 images by filtering out the users with less than 50 images.

Choosing the rank of factor matrices $r_U$, $r_I$ and $r_T$ in Tucker Decomposition model is not trivial. A practical option is to use ranks indicated by SVD on the unfolded matrices in each mode [34]. The tensor $\mathcal{Y}$ can be unfolded along different modes, leading to three new matrices $Y_U \in \mathbb{R}^{|\mathbb{U}|\times|\mathbb{I}||\mathbb{T}|}$, $Y_I \in \mathbb{R}^{|\mathbb{I}|\times|\mathbb{U}||\mathbb{T}|}$ and $Y_T \in \mathbb{R}^{|\mathbb{T}|\times|\mathbb{U}||\mathbb{I}|}$. In this way, $r_U$, $r_I$ and $r_T$ are chosen by preserving a certain percentage of singular values in the unfolded matrices. By fixing small values of $\alpha = 0.001$ and $\beta = 0.001$, we investigated the average F-score of tag refinement on NUS-WIDE-USER50 by tuning the percentage of the preserved energy from 50% to 95%. The result in Fig.7(a) indicates that 80% performs well on NUS-WIDE-USER50. By preserving 80% energy of the singular values, $r_U = 25$, $r_I = 105$ and $r_T = 18$.

The regularization terms $\alpha$ and $\beta$ control how much the tensor decomposition incorporates the information of affinity intra-relations. We keep $r_U = 25$, $r_I = 105$ and $r_T = 18$. Fig.7(b) shows the impacts of $\alpha$ and $\beta$ on the average F-score. $\alpha = 0.01$ and $\beta = 0.001$ achieves the best result. From the results, we can see that the performance is more sensitive to the regularization weights than to the rank numbers. The poor performances when $\alpha = 0$ or $\beta = 0$ confirm with the intuition that purely affinity constraints or *l*-1 norm constraints cannot generate good latent factors. For the remaining experiment, we select $r_U = 25$, $r_I = 105$, $r_T = 18$, $\alpha = 0.01$ and $\beta = 0.001$.

### C. Performance Comparison

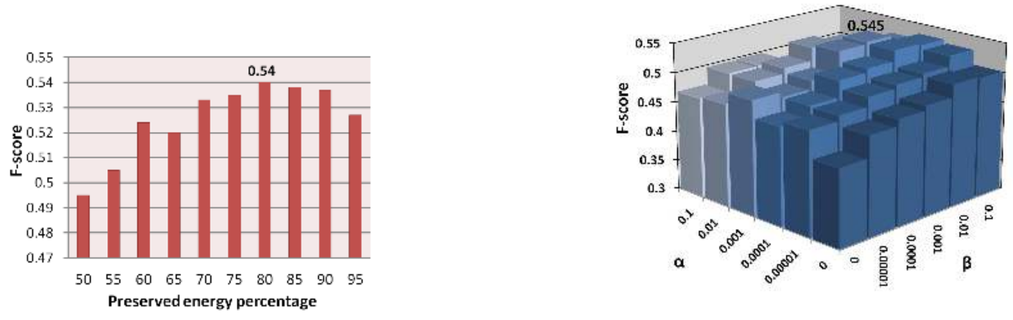
To compare the performances, five algorithms as well as the original tags are employed as the baselines:

- Original tagging (OT): the original user-generated tags.
- Random walk with restart (RWR): the tag refinement algorithm based on random walk [25].
- Tag refinement based on visual and semantic consistency (TRVSC, [9]).
- Multi-Edge graph (M-E Graph): a unified multi-edge graph framework for tag processing proposed in [28].
- Low-Rank approximation (LR): tag refinement based on low-rank approximation with content-tag prior and error sparsity [10].
- Multiple correlation Probabilistic Matrix Factorization (MPMF): the tag refinement algorithm by simultaneously modeling image-tag, tag-tag and image-image correlations into a factor analysis framework. [7].

In addition, we compared the performances of the proposed approach with four different settings: 1) TF without smoothness constraints, optimization under the *0/1 scheme* (TF_0/1), 2) TF with multi-correlation smoothness constraints, optimization under the *0/1 scheme* (MTF_0/1), 3) TF without smoothness constraints, optimization under the *ranking scheme* with negative set constructed as Eq.14 (TF_rank) and 4) TF with multi-correlation smoothness constraints, optimization under the *ranking scheme* with negative set constructed as Eq.10 (RMTF).

Table II lists the average performances for different tag refinement algorithms. It is shown that RWR fails on the noisy web data. One possible reason is that the model does not fully explore the image-image intra-relations. Both TRVSC and M-E Graph suffer from the high computation problem and the performances are limited on large-scale applications. As their methods are difficult to implement, the results of

[13] Due to link failures, the owner ID of some images is unavailable



(a) (b)

Fig. 7. Impact of parameters (a) rank numbers (b) $\alpha$ and $\beta$

TABLE II
AVERAGE PERFORMANCES OF DIFFERENT ALGORITHMS FOR TAG REFINEMENT

| | OT | RWR | TRVSC | M-E Graph | LR | MPMF | TF_0/1 | MTF_0/1 | TF_rank | RMTF |
|---|---|---|---|---|---|---|---|---|---|---|
| **F-score** | 0.477 | 0.475 | 0.490 | 0.530 | 0.523 | 0.521 | 0.515 | 0.542 | 0.531 | **0.571** |

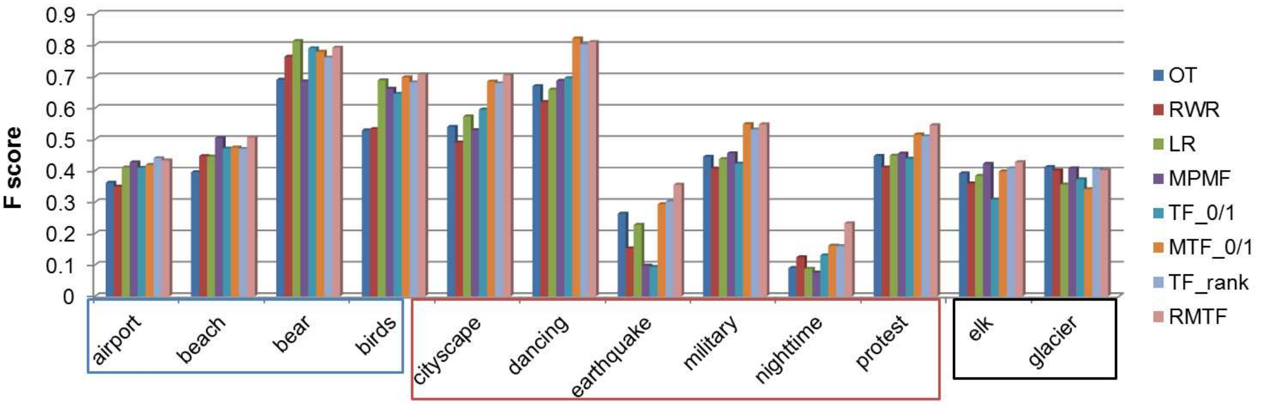


Fig. 8. F-score of a subset of the 81 tags for different algorithms

TRVSC and M-E Graph are taken from [28], which conducted tag refinement on a selected subset of NUS-WIDE. Their results on the whole NUS-WIDE dataset tend to decrease. Using factor analysis methods, MPMF and LR perform well on sparse dataset, which coincides with the authors' demonstration. For different settings of the proposed approach, RMTF, and MTF_0/1 are superior than other algorithms, showing the advantage of incorporating *user* information. Interpreting the tagging data based on the proposed *ranking scheme* instead of the conventional 0/1 *scheme*, RMTF is generally better than MTF_0/1. Without smoothness priors, TF_0/1 fails to preserve the affinity structures and achieves inferior results.

We note that TF_rank follows the same spirits as Rendle's works [30], [31] and was implemented to perform performance comparison with the proposed RMTF method. Consistent with the discussion in *section* IV.A that Rendle's works cannot fully account for the issues of missing tags and noisy tags, TF_rank obtains less improvement than the proposed RMTF. Actually, without consideration on the utilization of smoothness constraints, TF_rank is even inferior to MTF_0/1. In addition, according to the negative set selection strategy of TF_rank, the optimization algorithm needs to consider redundant pairs of training samples. It turns out that generally TF_rank achieves slower convergence speed than MTF_0/1 and RMTF.

The detailed performances for a representative subset of the 81 tags are provided in Fig.8. We can see that, for simple concepts like "airport", "beach", "bear" and "birds", our methods achieve a comparable, if not worse performance with the baselines. The reason is that images containing these concepts describe feasible and tangible objects, where image understanding can be effectively conducted by propagating visual similarities and only exploiting the *image-tag* relations. While, for more abstract and complex concepts like "cityscape", "earthquake", "military", "protest", existing methods focusing on utilizing image appearances and tag semantics fail and our methods show remarkable improvement thanks to the incorporation of *user* information. In addition, we also found that for those uncommon concepts like "elk" and "glacier", both the proposed methods and the baselines obtained no improvement and failed to perform image refinement. The failure of our methods may be due to the severe sparse user distribution on these concepts. Those uncommon concepts focalize to small groups, which make it difficult to propagate information between users.

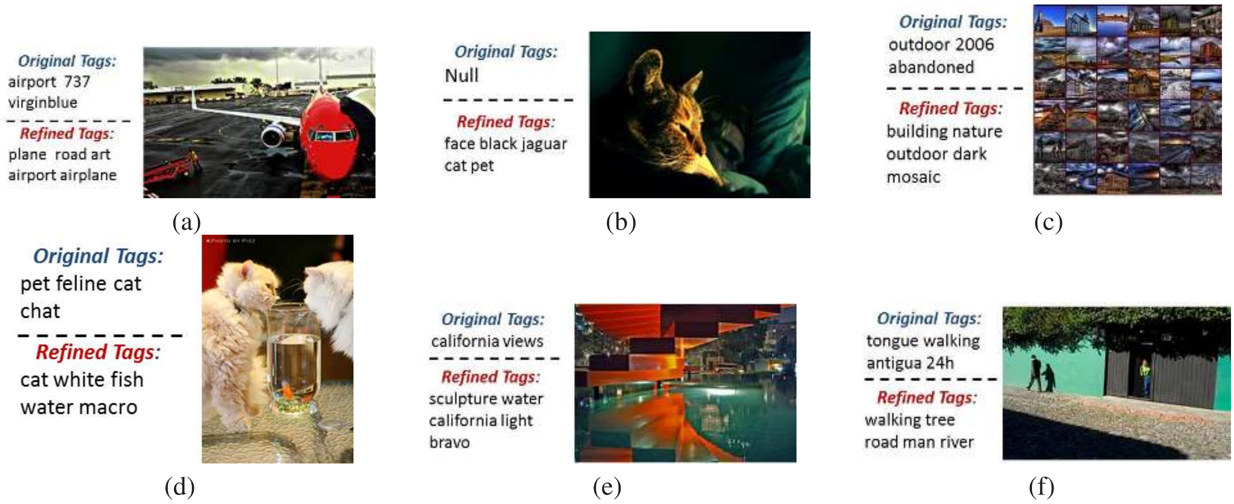


Fig. 9. Example of tag refinement results. For each image, the top 5 annotations are shown.

### D. Case Studies

We show some case studies in this subsection to demonstrate the effectiveness of RMTF. Fig.9 further illustrates the tag refinement results for some exemplary images by the proposed RMTF framework. For examples of Fig.9(c) and Fig.9(e), it is very hard to restore the relations between tags and images only from the visual appearance, since the images are very complex. With the aid of *user* information, it is observed that the tagger of Fig.9(c) also tagged "mosaic" and "building" to images and the tagger of Fig.9(d) is a "sculpture" fan. Therefore, the exploited semantic is propagated into the refined results. In the original tag set of Fig.9(a), only the tag "airport" is related to the image content. After tag refinement, the subjective tags are removed and the context-relevant tags, "airport", "road", and semantic-relevant tags "plane" are enriched through the proposed ranking-based optimization scheme. Fig.9(d)(f) further show this advantage. Moreover, Fig.9(b) demonstrates the capacity of the proposed framework on automatic image annotation. It can be seen that the experimental results validate our intuition that incorporation of *user* information with appropriate optimization scheme and smoothness constraints contributes to a better modeling of the tagging data and derives compact *image* and *tag* factor representations.

We have employed smoothness constraints into the optimization function to preserve the affinity structure in the low dimensional factor subspace. To show the effectiveness of smoothness constraints, we show in Table III and Table IV the five nearest tags and images for the selected tag and image, respectively. It is shown that RMTF succeeds to mine the semantic correlations among tags and images from the observed tagging data. Context and semantic relevant tags are close in the derived tag subspace, while in the image subspace, visual and sematic similar images are clustered together.

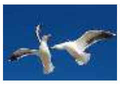
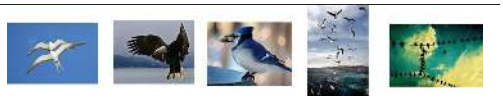
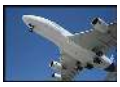
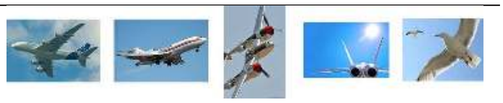
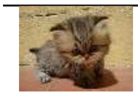
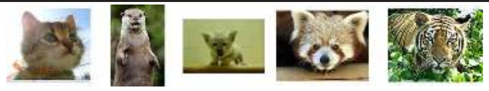
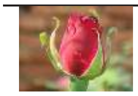
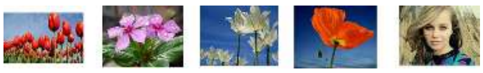
### E. Applications

In the tag refinement task, we employed the derived factor matrices to analyze the *image-tag* associations. As we model the social tagging data by taking into account all essential entities, *user*, *image* and *tag*, we can apply the model to many other real-world tasks.

TABLE III
FIVE NEAREST TAGS IN THE LEARNED TAG SUBSPACE FOR EACH OF THE FOUR SELECTED TAGS

| Selected Tag | Five Nearest Tags |
|---|---|
| cat | grass, animal, pet, dog, vacation |
| flower | blooms, butterfly, nature, spring, blossoms |
| airplane | aircraft, travel, planes, photographer, airport |
| buddhist | buddha, religion, buddhism, thailand, ancient |

TABLE IV
FIVE NEAREST IMAGES IN THE LEARNED IMAGE SUBSPACE FOR EACH OF THE FOUR SELECTED IMAGES

| Image | Five Nearest Images |
|---|---|
|  |  |
|  |  |
|  |  |
|  |  |

*1) Personalized image search:* In personalized image search, the returned image results depend on not only their relevances with the query keywords, but the relevances with the searchers. For our case, the associations between users and images can be estimated by measuring the *user-image* cross-space distances in the same spirits as Eq.5, which reflect the users' preferences and can be leveraged to re-rank the returned images.

An experiment is conducted. Following [35]'s evaluation framework, in the context of Flickr, the photos marked *Favorites* by the searcher are treated as the ground-truth. We

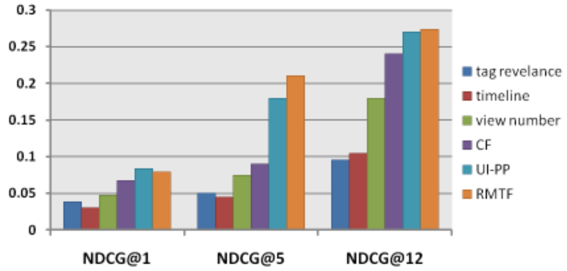


Fig. 10. Evaluation results for the personalized image search

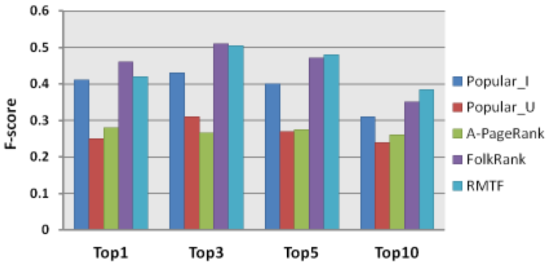


Fig. 11. Evaluation results for the personalized tag recommendation

chose 30 users who have the largest number of *Favorites* in the image collection as the searchers. 58 tags frequently appearing in their favorite images are selected as the queries. The metric of NDCG@k is utilized to evaluate the performance. For each query, we re-rank the top 50 (if there are) results by the tag-relevance, and average the evaluated scores over queries and searchers. The average results are demonstrated in Fig.10, where we compare with two personalized methods, user-based collaborative filtering (CF, [36]) and user interests-based preference prediction (UI-PP, [35]) and three non-personalized rules depending on relevance, view number and timelines. We can see that the three personalization methods outperform the non-personalized rules and RMTF achieves comparable performance with state-of-the-art.

*2) Personalized tag recommendation:* The goal of a personalized tag recommender is to predict tags for each user on a given web item (image, music, URL or publication). The reconstructed tensor $\hat{\mathcal{Y}}$ captures the ternary relationships between users, images and tags, where the value of $\hat{y}_{u_1,i_1,t_1}$ indicates the likelihood of user $u_1$ using tag $t_1$ to annotate image $i_1$. Therefore, the tags with the highest $\hat{y}_{u,i,t}$ can be recommended to user $u$ as the potential tags for item $i$.

We conducted the experiment on a small benchmark dataset from Bibsonomy,[14] which consists of 116 users, 412 tags and 361 items (publications). For each user, one post is randomly removed for evaluation. We averaged the F-scores in top-N recommended tags over users. Four personalized tag recommendation algorithms are performed as baselines: most popular tags by item (Popular_I), most popular tags by user (Popular_U), Adapted PageRank [37] and FolkRank [38]. Fig.11 illustrates the results. It is shown that with an increasing number of recommended tags, the F-score decreases less steeper for RMTF than other algorithms.

Note that we provide these two experiments to demonstrate the potentials of the proposed framework. As the focus of this paper is image tag refinement, we did not fully adapt RMTF to other applications. For example, typical methods of personalized tag recommendation (e.g., FolkRank) will consider the user and item bias, while in our implementations we did not explicitly consider this. With careful adaption to these applications, the performance of RMTF has the potential to improve.

The proposed RMTF can also be applied to other applications, e.g., **user profile construction** and **user recommendation**. It is believed that users express their individual interests through tags [39], thus the latent user interests can be understood by estimating the *user-tag* association. Besides exploring the interrelations, we can directly evaluate the intra-relations among users, images and tags in the corresponding subspaces. Users with similar feature representations can be recommended to each other to connect people with common interests and encourage people to contribute and share more content.

## VI. Conclusions

We have presented a ranking based multi-correlation factor analysis method that jointly models the *user*, *image* and *tag* factors. We argued that by exploiting the underlying structure of the photo sharing websites, our model is able to learn more semantics-specified image and tag descriptions from a corpus of social tagging data. The experimental results on collections from the photo sharing site Flickr show that our model performs well on the tag refinement task.

The potential applications and two simple experiments are also presented in the paper. It is an interesting issue to adapt the proposed RMTF to more related applications in the future. In addition, there exist different forms of metadata, such as descriptions, comments, and ratings. While we focus on tags in this paper, how to model other metadata for a overall understanding is also our future work.

## Appendix
## Proof of Theorem 2

*Proof:* For easier explanation, we rewrite the optimization function of Eq.21 into a general form:

$$\min_{\Theta\in\mathbb{X}} g(\Theta) \tag{23}$$

where $\Theta$ are the model parameters of $U, I, T, \mathcal{C}$ and denoted as $\Theta = (\theta_1, \theta_2, \theta_3, \theta_4)$, $\mathbb{X}$ is a Cartesian product of closed convex sets $\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3, \mathbb{X}_4$:

$$\mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2 \times \mathbb{X}_3 \times \mathbb{X}_4 \tag{24}$$

We assume that $\mathbb{X}_1$, $\mathbb{X}_2$, $\mathbb{X}_3$, $\mathbb{X}_4$ are closed convex subsets of $\mathbb{R}^{r_U\times r_I\times r_T}$, $\mathbb{R}^{|\mathbb{U}|\times r_U}$, $\mathbb{R}^{|\mathbb{I}|\times r_I}$, $\mathbb{R}^{|\mathbb{T}|\times r_T}$, respectively. Then the subproblems in Algorithm 1 can be formulated into a unique form:

$$\theta_i^{(t+1)} = \arg\min_{\theta_i\in\mathbb{X}_i} g(\theta_1^{(t+1)}, \cdots, \theta_{i-1}^{(t+1)}, \theta_i, \theta_{i+1}^{(t)}, \cdots, \theta_4^{(t)}) \tag{25}$$

According to *Theorem* 1, the minimum in Eq.25 is uniquely attained. In the following, we first prove the algorithm will

[14] http://www.bibsonomy.org/

converges to a limit point, and then show the limit point is a local optimum.

An auxiliary vector is introduced:

$$Z_i^{(t)} := (\theta_1^{(t+1)}, \cdots, \theta_i^{(t+1)}, \theta_{i+1}^{(t)}, \cdots, \theta_4^{(t)})$$

By Eq.25, we obtain

$$g(\Theta^{(t)}) \geq g(Z_1^{(t)}) \geq \cdots \geq g(Z_3^{(t)}) \geq g(\Theta^{(t+1)}), \forall t \quad (26)$$

Let $\bar{\Theta} = (\bar{\theta}_1, \cdots, \bar{\theta}_4)$ be a limit point the sequence $\{\Theta^{(t)}\}$. Since $\mathbb{X}$ is closed, $\bar{\Theta} \in \mathbb{X}$. Eq.26 implies sequence $\{g(\Theta^{(t)})\}$ converges to $g(\bar{\Theta})$.

Let $\{\Theta^{(t_j)} | j = 0, 1, \cdots\}$ be a subsequence of $\{\Theta^{(t)}\}$. We first show that $\{Z_1^{(t_j)} - \Theta^{(t_j)}\}$ converges to zero as $j \to \infty$. Assuming the contrary that $\{Z_1^{(t_j)} - \Theta^{(t_j)}\}$ does not converge to zero, we define $\gamma^{(t_j)} = ||Z_1^{(t_j)} - \Theta^{(t_j)}||$ and $\exists \hat{\gamma}, \gamma^{(t_j)} \geq \hat{\gamma}$. Let $\boldsymbol{s}^{(t_j)} = (Z_1^{(t_j)} - \Theta^{(t_j)})/\gamma^{(t_j)}$. Thus, $Z_1^{(t_j)} = \Theta^{(t_j)} + \gamma^{(t_j)}\boldsymbol{s}^{(t_j)}$, $\boldsymbol{s}_1^{(t_j)} = 1$ and $\boldsymbol{s}_{2,3,4}^{(t_j)} = 0$. Fix some $\epsilon \in [0, 1]$ with $0 \leq \epsilon\hat{\gamma} \leq \gamma^{(t_j)}$. Therefore,

$$g(Z_1^{(t_j)}) = g(\Theta^{(t_j)} + \gamma^{(t_j)}\boldsymbol{s}^{(t_j)}) \leq g(\Theta^{(t_j)} + \epsilon\hat{\gamma}\boldsymbol{s}^{(t_j)}) \leq g(\Theta^{(t_j)})$$

We assume $\lim_{j \to \infty} \boldsymbol{s}_1^{(t_j)} = \bar{\boldsymbol{s}}$ and take the limit of the above equation as $j \to \infty$, to obtain: $g(\bar{\Theta}) \leq g(\bar{\Theta} + \epsilon\hat{\gamma}\bar{\boldsymbol{s}}) \leq g(\bar{\Theta})$. We have

$$g(\bar{\Theta}) = g(\bar{\Theta} + \epsilon\hat{\gamma}\bar{\boldsymbol{s}}), \quad \forall \epsilon \in [0, 1]$$

Since $\hat{\gamma}\bar{\boldsymbol{s}} \neq 0$, this contradicts the fact that $g$ is uniquely minimized w.r.t. each subproblem. Therefore, we conclude that

$$\lim_{j \to 0} Z_1^{(t_j)} - \Theta^{(t_j)} = \mathbf{0}$$

From Eq.25, we have

$$g(Z_1^{(t_j)}) \leq g(\theta_1, \theta_2^{(t_j)}, \theta_3^{(t_j)}, \theta_4^{(t_j)}), \quad \forall \theta_1 \in \mathbb{X}_1$$

Taking the limit as $j \to \infty$, we obtain

$$g(\bar{\Theta}) \leq g(\theta_1, \bar{\theta}_2, \bar{\theta}_3, \bar{\theta}_4), \quad \forall \theta_1 \in \mathbb{X}_1 \quad (27)$$

Similar conclusions can be obtained for $\theta_2$, $\theta_3$ and $\theta_4$, and we conclude that $\bar{\Theta}$ minimizes $g$ over $\mathbb{X}$. Combining with the converge conclusion proved above, $g$ is guaranteed to converge to a stationary point. Because $g$ is not jointly convex w.r.t. $U$, $I$, $T$ and $\mathcal{C}$, the stationary point is a local optimum.

## References

[1] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, 2009.

[2] Dong Liu, Xian-Sheng Hua, and Hong-Jiang Zhang. Content-based tag processing for internet social images. *Multimedia Tools Appl.*, 51:723–738, January 2011.

[3] Yohan Jin, Latifur Khan, Lei Wang, and Mamoun Awad. Image annotations by combining multiple evidence & wordnet. In *ACM Multimedia*, pages 706–715, 2005.

[4] Changhu Wang, Feng Jing, Lei Zhang, and Hong-Jiang Zhang. Content-based image annotation refinement. In *CVPR*, 2007.

[5] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. Tag ranking. In *WWW*, pages 351–360, 2009.

[6] Hao Xu, Jingdong Wang, Xian-Sheng Hua, and Shipeng Li. Tag refinement by regularized lda. In *ACM Multimedia*, pages 573–576, 2009.

[7] Zechao Li, Jing Liu, Xiaobin Zhu, Tinglin Liu, and Hanqing Lu. Image annotation using multi-correlation probabilistic matrix factorization. In *ACM Multimedia*, pages 1187–1190, 2010.

[8] Lin Chen, Dong Xu, Ivor Wai-Hung Tsang, and Jiebo Luo. Tag-based web photo retrieval improved by batch mode re-tagging. In *CVPR*, pages 3440–3446, 2010.

[9] Dong Liu, Xian-Sheng Hua, Meng Wang, and Hong-Jiang Zhang. Image retagging. In *ACM Multimedia*, pages 491–500, 2010.

[10] Guangyu Zhu, Shuicheng Yan, and Yi Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM Multimedia*, pages 461–470, 2010.

[11] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[12] M. Alex O. Vasilescu and Demetri Terzopoulos. Multilinear subspace analysis of image ensembles. In *CVPR (2)*, pages 93–99, 2003.

[13] Xiaofei He, Deng Cai, and Partha Niyogi. Tensor subspace analysis. In *NIPS*, 2005.

[14] Xi Li, Weiming Hu, Zhongfei Zhang, Xiaoqin Zhang, and Guan Luo. Robust visual tracking based on incremental tensor subspace learning. In *ICCV*, pages 1–8, 2007.

[15] Jing Wen, Xinbo Gao, Yuan Yuan, Dacheng Tao, and Jie Li. Incremental tensor biased discriminant analysis: A new color-based visual tracking method. *Neurocomputing*, 73(4-6):827–839, 2010.

[16] Tamara G. Kolda and Jimeng Sun. Scalable tensor decompositions for multi-aspect data mining. In *ICDM*, pages 363–372, 2008.

[17] Thomas Franz, Antje Schultz, Sergej Sizov, and Steffen Staab. Triplerank: Ranking semantic web data by tensor decomposition. In *International Semantic Web Conference*, pages 213–228, 2009.

[18] P. Symeonidis, M. M. Ruxanda, A. Nanopoulos, and Y. Manolopoulos, "Ternary semantic analysis of social tags for personalized music recommendation," in *ISMIR*, 2008, pp. 219–224.

[19] C. Wei, W. Hsu, and M.-L. Lee, "A unified framework for recommendations based on quaternary semantic analysis," in *SIGIR*, 2011, pp. 1023–1032.

[20] Yu-Ru Lin, Jimeng Sun, Paul Castro, Ravi B. Konuru, Hari Sundaram, and Aisling Kelliher. Metafac: community discovery via relational hypergraph factorization. In *KDD*, pages 527–536, 2009.

[21] Tong Zhang and Gene H. Golub. Rank-one approximation to high order tensors. *SIAM J. Matrix Anal. Appl.*, 23:534–550, February 2001.

[22] Tamara G. Kolda, Brett W. Bader, and Joseph P. Kenny. Higher-order web link analysis using multilinear algebra. In *ICDM*, pages 242–249, 2005.

[23] Yun Chi and Shenghuo Zhu. Facetcube: a framework of incorporating prior knowledge into non-negative tensor factorization. In *CIKM*, pages 569–578, 2010.

[24] Ilya Sutskever, Ruslan Salakhutdinov, and Joshua Tenenbaum. Modelling relational data using bayesian clustered tensor factorization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1821–1828, 2009.

[25] Changhu Wang, Feng Jing, Lei Zhang, and HongJiang Zhang. Image annotation refinement using random walk with restarts. In *ACM Multimedia*, pages 647–650, 2006.

[26] Jing Liu, Bin Wang, Mingjing Li, Zhiwei Li, Wei-Ying Ma, Hanqing Lu, and Songde Ma. Dual cross-media relevance model for image annotation. In *ACM Multimedia*, pages 605–614, 2007.

[27] Lexing Xie, Apostol Natsev, Matthew L. Hill, John R. Smith, and Alex Phillips. The accuracy and value of machine-generated image tags: design and user evaluation of an end-to-end image tagging system. In *CIVR*, pages 58–65, 2010.

[28] Dong Liu, Shuicheng Yan, Yong Rui, and Hong-Jiang Zhang. Unified tag analysis with multi-edge graph. In *ACM Multimedia*, pages 25–34, 2010.

[29] Yoshua Bengio, Jean-François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *NIPS*, 2003.

[30] S. Rendle, L. B. Marinho, A. Nanopoulos, and L. Schmidt-Thieme, "Learning optimal ranking with tensor factorization for tag recommendation," in *KDD*, 2009, pp. 727–736.

[31] Steffen Rendle and Lars Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *WSDM*, pages 81–90, 2010.

[32] Wu-Jun Li and Dit-Yan Yeung. Relation regularized matrix factorization. In *IJCAI*, pages 1126–1131, 2009.

[33] Yunhong Zhou, Dennis M. Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. In *AAIM*, pages 337–348, 2008.

[34] Evrim Acar and Bülent Yener. Unsupervised multiway data analysis: A literature survey. *IEEE Trans. Knowl. Data Eng.*, 21(1):6–20, 2009.

[35] Dongyuan Lu and Qiudan Li. Personalized search on flickr based on searcher's preference prediction. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, 2011.

[36] U. Rohini and Vamshi Ambati. A collaborative filtering based re-ranking strategy for search in digital libraries. In *ICADL*, pages 194–203, 2005.

[37] Robert Jäschke, Leandro Balby Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in folksonomies. In *PKDD*, pages 506–514, 2007.

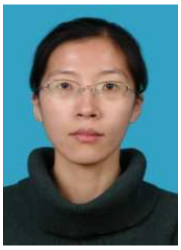
[38] Robert Jäschke, Leandro Balby Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in social bookmarking systems. *AI Commun.*, 21(4):231–247, 2008.

[39] Ching man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. A study of user profile generation from folksonomies. In *SWKM*, 2008.

**Jitao Sang** received the B.E. degree from the South-East University, Nanjing, China, in 2007. He is currently pursuing the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing.

In 2010 and 2011, he was an intern student in the China-Singapore Institute of Digital Media (CSIDM) and Microsoft Research Asia (MSRA), respectively. His current research interests include multimedia content analysis, social media mining, computer vision and pattern recognition.

**Changsheng Xu** (M'97-SM'99) is Professor in National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences and Executive Director of China-Singapore Institute of Digital Media. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition and computer vision. He has hold 30 granted/pending patents and published over 200 refereed research papers in these areas. Dr. Xu is an Associate Editor of ACM Transactions on Multimedia Computing, Communications and Applications and ACM/Springer Multimedia Systems Journal. He served as Program Chair of ACM Multimedia 2009. He has served as associate editor, guest editor, general chair, program chair, area/track chair, special session organizer, session chair and TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences and workshops.



**Jing Liu** (M08) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2008.

She is an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her current research interests include machine learning, image content analysis and classification, multimedia information indexing and retrieval, and others.