Exploiting User Information for Image Tag Refinement *

Jitao Sang[1,2], Jing Liu[1,2], Changsheng Xu[1,2]
[1]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
[2]China-Singapore Institute of Digital Media, Singapore, 119613, Singapore
{jtsang, jliu, csxu}@nlpr.ia.ac.cn

ABSTRACT

Photo sharing websites allow users to describe images with freely chosen tags. The user-generated tags not only facilitate the users in sharing and organizing images, but also provide large scale meaningful data for image retrieval and management. Extensive studies on improving the quality of user-generated tags for tag-based applications focused on exploiting the *image-tag*, *image-image* and *tag-tag* binary relationships. Considering that *user* is the originator of the tagging activity and *user* involves with *image* and *tag* in many aspects, in this paper we tackle the problem of tag refinement by leveraging *user* information. We propose a Tensor Decomposition framework to jointly model the ternary *user-image-tag* interrelation and respective intra-relations. The users, images and tags are represented in the corresponding latent subspaces. For a given image, the tags with the highest cross-space associations are reserved as the final annotation. The proposed method is validated on a large-scale real-world dataset.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.8 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Theory, Experimentation

Keywords

Tag Refinement, Social Images, Factor Analysis

1. 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, Picasa, Zooomr and Pinterest. 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. Typically, in a photo sharing website (Flickr as example), 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.

Figure 1: Example images from Flickr and their associated tags and taggers

Obviously, given such a large-scale web dataset, noisy and missing tags are inevitable, which limits the performance of social tag-based retrieval system. Therefore, the tag refinement to denoise and enrich tags for images is desired to tackle this problem. Existing efforts on tag refinement [1, 2, 3, 4, 5, 6, 7, 8] 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.

Users are the originator of the tagging activity and they involves 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 a simple example to explain this observation. As shown in Fig.1, the both images are tagged with "jaguar" by the two users (indicated by user ID), 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'. Note that it is not necessary to explicitly know the users' interests or profiles. What we are interested in is the variations in individual user's tagging patterns and preferences.

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. With the user factor incorporated, the image and

*Area chair: Lexing Xie

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

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 *Multi-correlation Regularized Tensor Factorization* (MRTF) is proposed to tackle the tag refinement task. We utilize tensor factorization to jointly model the multiple factors. To alleviate the sparsity problem, 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 tensor factorization.
- We propose the MRTF model to extract the latent factor representations. The sparsity problem is alleviated by imposing the smoothness constraints.
- The proposed framework is evaluated on a large-scale dataset and the advantage of incorporating user information is validated.

2. 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 involve with 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 regularized tensor factorization method, 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., $\mathbb{U}, \mathbb{I}, \mathbb{T}$).

2.1 Multi-correlation Regularized 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. The left image in Fig.1 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 $\mathcal{Y} \in \mathbb{R}^{|\mathbb{U}| \times |\mathbb{I}| \times |\mathbb{T}|}$,

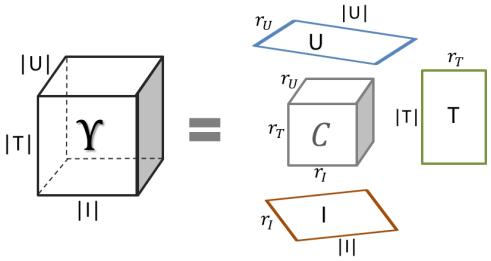

Figure 2: Tucker decomposition: the tensor $\mathcal{Y}$ is constructed by multiplying three factor matrices U, I, T to a small core tensor $\mathcal{C}$.

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.

To jointly model the three factors of *user*, *image* and *tag*, we employ a tensor factorization model, Tucker Decomposition for the latent factors inference. In Tucker Decomposition, the tagging data $\mathcal{Y}$ are estimated by three low rank matrices and one core tensor (see Fig.2):

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

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 this is called *rank-*(r_U, r_I, r_T) Tucker decomposition. An intuitive interpretation of Eq.2 is that the tagging data depends not only on how similar an image's visual features and tags' 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 decomposition 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 $\hat{y}_{\tilde{u},\tilde{i},\tilde{t}} = \mathcal{C} \times_u \boldsymbol{u}_{\tilde{u}} \times_i \boldsymbol{i}_{\tilde{i}} \times_t \boldsymbol{t}_{\tilde{t}}$.

In addition to the ternary interrelations, we also collect multiple intra-relations among users, images and tags. These intra-relations constitute the 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. We assume that two items 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 the following, 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 indicate the users' interests and backgrounds. Therefore, we measure the affinity relationship between user u_m and u_n using the co-occurrence of their joined groups:

$$W^U_{m,n} = \frac{n(u_m, u_n)}{n(u_m) + n(u_n)} \tag{4}$$

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 [8, 7], 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:

$$W^I_{m,n} = e^{-||\boldsymbol{d}_m - \boldsymbol{d}_n||^2/\sigma_I^2} \tag{5}$$

where σ_I is set as the median value of the elements in W^I.

Tag affinity graph W^T. We build the tag affinity graph based on the tag co-occurrance relevance. The relevance of tag t_m and t_n is simply encoded by their weighted co-occurrence in the image collection:

$$W^T_{m,n} = \frac{n(t_m, t_n)}{n(t_m) + n(t_n)} \tag{6}$$

Note we have no rigid requirements for how to build the affinity graphs and other intra-relation measurements can also be explored.

The affinity graphs can be 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 and the image factor matrix I as example, the regularization term is:

$$\sum_{m=1}^{|\mathbb{I}|}\sum_{n=1}^{|\mathbb{I}|} W^I_{m,n}||\boldsymbol{i_m} - \boldsymbol{i_n}||^2 \tag{7}$$

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:

$$\begin{aligned} l &= \sum_{m=1}^{|\mathbb{I}|}\sum_{n=1}^{|\mathbb{I}|} W^I_{m,n}||\boldsymbol{i_m} - \boldsymbol{i_n}||^2 \\ &= \sum_{d=1}^{r_I}\sum_{m=1}^{|\mathbb{I}|}\sum_{n=1}^{|\mathbb{I}|} W^I_{m,n}(i_{m,d} - i_{n,d})^2 = tr(I^\top L_I I) \end{aligned} \tag{8}$$

where $tr(\cdot)$ denotes the trace of a matrix and L_I is the Laplacian matrix for the image affinity matrix W^I. We can build similar regularization terms for the user and tag factors. Combining with Eq.3, we obtain the following overall objective function:

$$\begin{aligned} \min_{U,I,T,\mathcal{C}} g = &\sum_{(\tilde{u},\tilde{i},\tilde{t})\in|\mathbb{U}|\times|\mathbb{I}|\times|\mathbb{T}|} (\hat{y}_{\tilde{u},\tilde{i},\tilde{t}} - y_{\tilde{u},\tilde{i},\tilde{t}})^2 \\ &+ \alpha(tr(U^\top L_U U) + tr(I^\top L_I I) + tr(T^\top L_T T)) \\ &+ \beta(||\mathbb{U}||^2 + ||\mathbb{I}||^2 + ||\mathbb{T}||^2) \end{aligned} \tag{9}$$

where $||\mathbb{U}||^2 + ||\mathbb{I}||^2 + ||\mathbb{T}||^2$ is l-1 regularization term to penalize large parameters, α and β are weights controlling the strength of corresponding constraints. Obviously, directly optimizing Eq.9 is infeasible and we use an iterative optimization algorithm. We propose an alternating learning algorithm (ALA) with gradient descent to learn the latent factors by iteratively optimizing each subproblems.

Table 1: 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

2.2 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. Each row of the factor matrices corresponds to one item (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 $\mathcal{T}^{UI} := \mathcal{C} \times_t T$, 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:

$$X^{IT} = I \cdot (\mathcal{C} \times_t T \times_u \mathbf{1}^\top_{r_U}) \tag{10}$$

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

$$Top(i, K) = \max_{t\in\mathbb{T}}{}^K X^{IT}_{i:} \tag{11}$$

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

3. EXPERIMENTS

3.1 Data Set

We perform the experiments of social tag refinement on the large-scale web image dataset, NUS-WIDE [9]. 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[1]. 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 1 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:

$$Fscore = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

3.2 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 (RWRW): the tag refinement algorithm based on random walk [10].
- Tag refinement based on visual and semantic consistency (TRVSC, [7]).

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

Table 2: Average performances of different algorithms for tag refinement

	OT	RWRW	TRVSC	M-E Graph	LR	MPMF	TF	MRTF
F-score	0.477	0.475	0.49	0.53	0.523	0.521	0.515	0.539

Figure 3: Example of tag refinement results. For each image, the top 5 annotations are shown.

- Multi-Edge graph (M-E Graph): a unified multi-edge graph framework for tag processing proposed in [11].
- Low-Rank approximation (LR): tag refinement based on low-rank approximation with content-tag prior and error sparsity [8].
- Multiple correlation Probabilistic Matrix Factorization (MPMF): the tag refinement algorithm by simultaneously modeling image-tag, tag-tag and tag-tag correlations into a factor analysis framework. [5].

In addition, we compared the performances of the proposed approach with different settings: 1) TF without smoothness constraints (TF) 2) TF with multi-correlation smoothness constraints (MRTF_0/1). Table 2 lists the average performances for different tag refinement algorithms.

It is shown that RWRW fails on the noisy web data. One possible reason is that the model does not fully explore the image-image intra-relations. The results of TRVSC and M-E Graph are taken from [11], which conducted tag refinement on a smaller subset of NUS-WIDE. Both TRVSC and M-E Graph suffer from the high computation problem and the performances are limited on large-scale applications. Using factor analysis methods, MPMF and LR perform well on sparse dataset, which coincides with the authors' demonstration. MRTF is superior than the other methods, showing the advantage of incorporating *user* information. Without smoothness priors, TF fails to preserve the affinity structures and achieves inferior results. Fig.3 further shows the tag refinement results for some exemplary images by the proposed MRTF framework. The experimental results validate our intuition that incorporation of *user* information with appropriate smoothness constraints contribute to a better modeling of the tagging data and derive compact *image* and *tag* factor representations.

4. CONCLUSIONS

We have presented a multi-correlation regularized 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.

In the future, 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 one of our future works.

5. ACKOWNLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (Grant No. 60903146, 90920303), 973 Program (Project No. 2010CB327905) and Microsoft Research Asia UR Project.

6. REFERENCES

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

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

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

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

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