Chapter 2
User-Perceptive Multimedia Content Analysis

Abstract Typical social multimedia services allow users as uploaders, viewers, taggers, and commenters to interact and collaborate with each other in a communication dialog. The wisdom of crowds provides a huge resource for understanding social multimedia content. In this chapter, we explicitly model user interaction in the tag generation process and propose a regularized tensor factorization solution to refine the ternary correlations among user, image, and tag. While the traditional social tag analysis work focus on analyzing the image-tag binary correlation, taking user factor into consideration shows superior performance in image tag refinement task.

2.1 Introduction

Multimedia content analysis is the first step for most traditional multimedia computing tasks. In social multimedia computing, current social multimedia platforms allow users interacting with multimedia through uploading, annotating, commenting, and interacting with each other through social dialogs. These interactions capture what user perceive the multimedia content, and can be exploited toward multimedia content analysis, e.g., user-contributed picture tags indicate user-perceived visual semantics, user browsing behaviors, such as pause, fast-forward, indicate video structure information. In this section, we will review existing work in user-perceptive multimedia content analysis based on the exploited interactions.

The idea of exploiting the crowd wisdom from user interaction for multimedia content analysis has been realized into several popular systems. The best example goes to the ESP game [38], which is designed to make collaboratively people label images as a side-effect of playing a game. The derived image labels can be used as training samples for image annotation and to help improve image search on the Web. Another example is Waze, which is another system based on user collaboration. Acquired by Google in 2013, Waze is a free turn-by-turn GPS application for mobile phones that uses crowdsourcing to provide routing and real-time traffic updates. Other successful systems include Wikipedia, the world's largest free encyclopedia that is written collaboratively by anonymous Internet contributors, and Facebook Translations, by

J. Sang, *User-centric Social Multimedia Computing*, Springer Theses,
DOI 10.1007/978-3-662-44671-3_2

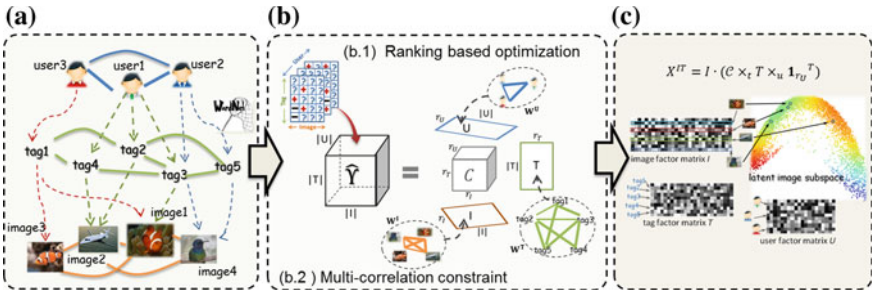


Fig. 2.1 The proposed solution framework. **a** Data Collection. **b** RMTF. **c** Tag Refinement. ©[2012] IEEE. Reprinted, with permission, from Ref. [34]

which over 3 million users voluntarily help translate Facebook webpages into 60 different languages.

With the popularity of Web 2.0, there are explosive photo sharing websites with large-scale image collections available online, such as Flickr, Pinterest, Instagram, Picasa, etc. These Web 2.0 websites allow users as owners, taggers, or commenters for their contributed images, leading to a huge amount of social images with user-contributed tags. Obviously, given such a large-scale web dataset, noisy and missing tags are inevitable, which limit the performance of social tag-based retrieval system [5, 22]. Therefore, the tag refinement to denoise and enrich tags for images is desired to tackle this problem. Existing efforts on tag refinement [4, 16, 19–21, 40, 42, 47] 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.

The goal of this chapter is to introduce the user factor into social image tag analysis tasks, and improve the underlying associations between the images and tags from the observed raw tagging data. To this end, we address the tag refinement problem 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 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 Multicorrelation Tensor Factorization* (RMTF) is proposed to tackle the tag refinement task. The framework is illustrated in Fig. 2.1.[1] It contains three components: 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 intrarelations are collected. 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 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

[1] We show a running example consisting of three users, five tags, and four images in Fig. 2.1a.

tags are considered in a conservative filtering strategy by exploiting the tag correlation on context and semantics. Besides, the multiple intrarelations 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. Most of the work in this chapter has been published in [33, 34].

2.2 Related Work

2.2.1 Multimedia Content Analysis

The idea of exploiting user perception has been realized in many social multimedia content analysis tasks. In this subsection, as summarized in Table 2.1, we will review related work based on the exploited user interactions.

The first type of user interaction is the metadata associated with social multimedia content, e.g., descriptions, tags. Such metadata provides a natural context for multimedia content analysis, which helps reducing the semantic gap between low-level multimodal features and the high-level semantics. Since the user-contributed tags are generally noisy, ambiguous, incomplete or subjective, one critical research topic in exploiting the metadata is tag processing. Typical tag processing problems include tag ranking [21], tag refinement [47], tag-to-region [25], etc.[2] After tag processing, the second research line is devoted to semantic ontology construction from the processed

Table 2.1 The categorization of the work on user-perceptive multimedia content analysis according to the exploited user interactions

Exploited user interaction	Related work
Metadata	Tag processing [21, 25, 47]
	Semantic ontology construction [13, 29]
	Multimedia check-in mining [6, 44]
	Tag generation modeling [26]
Usage data	Browsing behavior [3, 28, 45]
	Social endorsements [15, 17]
	Comments [7, 8, 11, 12, 30, 36, 43]
User–user interaction	Undirected relation [14]
	Directed relation [9]
	Hybrid [10, 37]

[2] Note in most tag processing work, while tag is contributed by users, user factor is not explicitly considered. We will discuss the difference between our work in this chapter and the existing tag process work in next subsection.

tags. In [13], the authors studied the problem of constructing tag hierarchies from social tagging data, and investigated into the usefulness of such tag hierarchies in supporting efficient navigation from an information retrieval perspective. In [29], the issues of sparse, shallow, ambiguous, noisy, and inconsistent are considered when exploiting the structured metadata to folksonomy learning. A relational clustering solution is proposed. The third line is on exploiting the geographical metadata, e.g., geographical tag or check-in record. Ye et al. [44] develops a semantic annotation technique for location-based social networks to automatically annotate all places with category tags. In the Livehoods Project [6] exploits the check-in records collected from a location-based social network to study the structure and composition of a city, which requires long hours of observation and interviews in traditional means. Another interesting line is on interpreting and modeling the tag creation process from a generative perspective. One inspiring work is [26], where a new probabilistic generative model is proposed to simulate the generation process of social annotations from both resource topics and user perspectives.

Another important type of user interaction is user usage data recorded during social multimedia activities. Users unintentionally embed their understanding of the multimedia content in their interaction. Related works are reviewed along three lines. (1) Exploiting the browsing behaviors. An early work is conducted to utilize video browsing log, e.g., pause, forward, to understand the video semantic structure, and with applications in video browsing and summarization [45]. In [3], the authors investigated into the impact of previous views to the future popularity of YouTube videos, and found a "rich-get-richer" phenomenon. Inspired by the phenomenon, [28] presents a solution to employing the early view patterns to predict the future popularity, with potential applications in targeted advertising, effective search and recommendation services. (2) Exploiting social endorsements. Many social multimedia networking platform allows users to endorse entities that they find appealing. Reference [17] exploits the typical social endorsement activities, "favoriting" photos in Flickr, to extract relevant and descriptive entity semantics. In [15], a system called "LinkMiner" is developed to employ Facebook "Like" to understand use interest and estimate representativeness and influence of objects. (3) Exploiting comments. Reference [7] conducts a pioneer work on examining the motivations that users participate into YouTube video commenting conversations. In [30], the authors introduced an interesting work to exploit comments to analyze the cross-media similarity between textual and video items. Instead of focusing on cross-modal association analysis, the associated comments are employed as bridges for cross-media analysis and retrieval. Moreover, comments have also been exploited in other multimedia content analysis tasks, including inferring video semantics [8, 11], estimating the mood of music or video [36, 43], predicting item popularity [12], and so on.

Besides the users' direct interaction with social multimedia objects, very recently, the interactions between users are also exploited to address the social multimedia content analysis tasks. The social interactions between users, i.e., social relations, can be categorized into undirected and directed relations. In [14], the authors has explored the utilization of undirected social relations to facilitate sentiment analysis in the context of microblogging. A social science theory called "emotion contagion" is

utilized, which assumes that people tend to catch others emotions as a consequence of facial, vocal, and postural feedback. In [9], the directed social relations, i.e., "contact" in Flickr, is exploited to estimate the photo popularity and influence at topic-level. Hypergraph is constructed to represent user, image and the heterogeneous relations between users and images. Some studies integrate the metadata, usage data and interactions between users towards social multimedia analysis. In [10], the authors proposed to exploit the heterogeneous information like users' tagging behaviors, social networks, tag semantics and item profiles to alleviate the cold start problem in recommender system. In [37], a hybrid solution is presented to identify the geographic location of web videos. Various available sources of information, e.g., user's upload history, the social network, video tagging, are exploited with a divide & conquer strategy.

2.2.2 Social Image Tag Refinement

The literatures [22, 47] provide good surveys for the research work on image tag refinement. Along the three basic elements in photo tagging behaviors, i.e., image, tag, and user, we characterize the related work according to the elements they leveraged.

As a pioneer work, Jin et al. [16] employed WordNet to estimate the semantic correlations among the annotated tags and remove weakly correlated ones. The work of [39] performs belief propagation among tags within the random walk with restart framework to refine the imprecise original annotations. In [42], 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 [24], 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. [21] proposed to rank the image tags according to their relevance w.r.t. the associated images by modeling tag similarity and image similarity. In [20], 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. [41], 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. [23] proposed a multiedge 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 chapter is [19, 47], which solves the tag refinement problem through low-rank matrix approximation. Zhu et al. [47] considered the tagging characteristics from the view of low-rank, error sparsity, content consistency and tag correlation. In [19], 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 chapter, 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.

2.3 Methods for Social Image Tag Refinement

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., $\mathscr{Y}$), 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., $\mathbb{U}, \mathbb{I}, \mathbb{T}$).

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 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 $\mathscr{Y} \in \mathbb{R}^{|\mathbb{U}| \times |\mathbb{I}| \times |\mathbb{T}|}$, which is defined as:

$$y_{u,i,t} = \begin{cases} 1 & \text{if } (u, i, t) \in \mathbb{O} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

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 the general tensor factorization model, Tucker Decomposition for the latent factor inference. In Tucker Decomposition, the tagging data $\mathscr{Y}$ are estimated by three low-rank matrices and one core tensor:

$$\hat{\mathscr{Y}} := \mathscr{C} \times_u U \times_i I \times_t T \tag{2.2}$$

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 $\mathscr{C} \in \mathbb{R}^{r_U \times r_I \times r_T}$ contains the interactions between the different factors. The ranks of decomposed factors are denoted by r_U, r_I, r_T and Eq. (2.2) is called *rank*-(r_U, r_I, r_T) Tucker decomposition. An intuitive interpretation of Eq. (2.2) is that the tagging data depends not only on how similar an image's visual features and tag's semantics are, but also on how much these features/semantics match with the users' preferences.

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

$$\min_{U,I,T,\mathscr{C}} \sum_{(\tilde{u},\tilde{i},\tilde{t}) \in |\mathbb{U}| \times |\mathbb{I}| \times |\mathbb{T}|} (\hat{y}_{\tilde{u},\tilde{i},\tilde{t}} - y_{\tilde{u},\tilde{i},\tilde{t}})^2 \tag{2.3}$$

where $\hat{y}_{\tilde{u},\tilde{i},\tilde{t}} = \mathscr{C} \times_u \boldsymbol{u}_{\tilde{u}} \times_i \boldsymbol{i}_{\tilde{i}} \times_t \boldsymbol{t}_{\tilde{t}}$. 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 chapter, we propose RMTF for factor inference, which is detailed in Sect. 2.3.1.

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 object (user, image or tag). The core tensor $\mathscr{C}$ defines a multilinear 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

$$\mathscr{T}^{UI} := \mathscr{C} \times_t T \tag{2.4}$$

then $\mathscr{T}^{UI} \in \mathbb{R}^{r_U \times r_I \times |\mathbb{T}|}$ 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 $\mathscr{T}^{UI}$ 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[3]:

$$X^{IT} = I \cdot (\mathscr{T}^{UI} \times_u \mathbf{1}_{r_U}^{\top}) \tag{2.5}$$

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

[3] 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 [2].

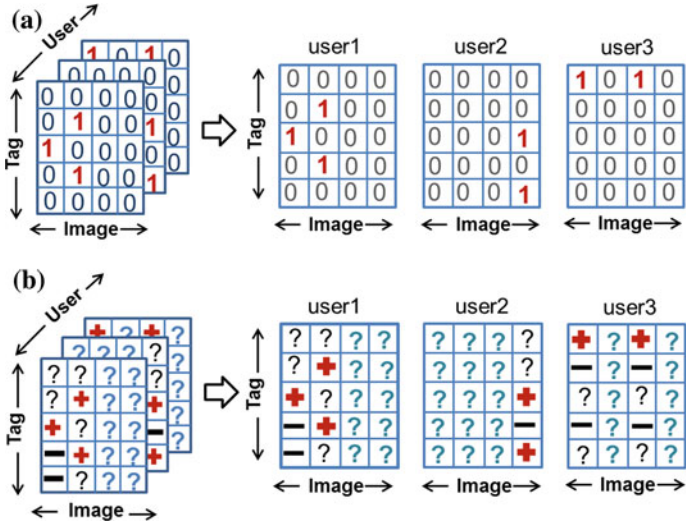


Fig. 2.2 Tagging data interpretation. **a** 0/1 scheme, **b** ranking scheme. ©[2012] IEEE. Reprinted, with permission, from Ref. [34]

$$\text{Top}(i, K) = \max_{t \in \mathbb{T}}{}^{K} X_{i:}^{IT} \tag{2.6}$$

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

2.3.1 Ranking-Based Optimization Scheme

Traditional factorization models [19, 47] 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 in Fig. 2.2a:

- First, 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.[4] Maybe he or she does not want to annotate the image or has no chance to see the image.
- Secondly, *user1* annotates *image1* with *tag3* only. 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. (2.3), the learning process tries to predict 0 for both cases, which is apparently unreasonable. To address the above

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

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^-.

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 $\mathbb{P}_{\mathbb{O}}$:

$$\mathbb{P}_{\mathbb{O}} = \left\{(u, i)|\exists t \in \mathbb{T}, y_{u,i,t} = 1\right\} \tag{2.7}$$

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

$$\mathbb{M} = \left\{(u, i, t)|(u, i) \notin \mathbb{P}_{\mathbb{O}}\right\} \tag{2.8}$$

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. 2.2b).

For the training pair determination, we consider two characteristics of the user tagging behaviors. On one hand, some concepts may be 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," and "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 intrarelations.[5] 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 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. 2.2b):

$$\mathbb{T}^+_{u,i} = \left\{t|(u, i) \in \mathbb{P}_{\mathbb{O}} \wedge y_{u,i,t} = 1\right\} \tag{2.9}$$

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

The negative tag set is constituted as:

$$\mathbb{T}^-_{u,i} = \left\{ t | (u,i) \in \mathbb{P}_{\mathbb{O}} \wedge y_{u,i,t} \neq 1 \wedge t \notin \mathbb{N}_{\mathbb{T}^+_{u,i}} \right\} \tag{2.10}$$

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 *tag*1 and *tag*2 are relevant to *tag*3. The final tagging data representation for the running example is illustrated in Fig. 2.2b. 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:

$$\hat{y}_{u,i,t_1} > \hat{y}_{u,i,t_2} \Leftrightarrow t_1 \in \mathbb{T}^+_{u,i} \wedge t_2 \in \mathbb{T}^-_{u,i} \tag{2.11}$$

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

$$\min_{U,I,T,\mathscr{C}} \sum_{(\tilde{u},\tilde{i}) \in \mathbb{P}_{\mathbb{O}}} \Big(\sum_{t^+ \in \mathbb{T}^+_{\tilde{u},\tilde{i}}} \sum_{t^- \in \mathbb{T}^-_{\tilde{u},\tilde{i}}} f(\hat{y}_{\tilde{u},\tilde{i},t^-} - \hat{y}_{\tilde{u},\tilde{i},t^+}) \Big) \tag{2.12}$$

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:

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

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 $(\tilde{u}, \tilde{i})$:

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

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

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

$\mathbf{p}$ 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:

$$f(\mathbf{p}^{\top}) \times \mathbf{1}_N \tag{2.13}$$

Note that the work in [31, 32] 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 are not annotated by *user u* to *image i* will be treated as negative tags, and the corresponding negative set is:

$$\mathbb{T}^{-}_{u,i} = \left\{ t | (u, i) \in \mathbb{P}_{\mathbb{O}} \wedge y_{u,i,t} \neq 1 \right\} \tag{2.14}$$

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 et al. employed l-1 norm for regularization, while in the proposed RMTF, additional multiple intrarelations are utilized as the smoothness constraints, which are detailed in the following subsection.

2.3.2 Multicorrelation Smoothness Constraints

In addition to the ternary interrelations, we also collect multiple intrarelations among users, images, and tags. These intrarelations constitute the user, image, and 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 intrarelations 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 cooccurrence of their joined groups:

$$W^U_{m,n} = \frac{n(u_m, u_n)}{n(u_m) + n(u_n)} \tag{2.15}$$

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 [20, 47], 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:

$$W^I_{m,n} = e^{-||\boldsymbol{d}_m - \boldsymbol{d}_n||^2/\sigma_I^2} \tag{2.16}$$

where σ_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 cooccurrence in the image collection:

$$t^c_{m,n} = \frac{n(t_m, t_n)}{n(t_m) + n(t_n)} \tag{2.17}$$

For tag semantic relevance, we follow Liu et al. [20] approach and estimate the semantic relevance between tag t_m and t_n based on their WordNet distance:

$$t^s_{m,n} = \frac{2 \cdot IC(lcs(t_m, t_n))}{IC(t_m) + IC(t_n)} \tag{2.18}$$

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:

$$W^T_{m,n} = \lambda_c t^c_{m,n} + \lambda_s t^s_{m,n} \tag{2.19}$$

where $\lambda_c + \lambda_s = 1$, λ_c and λ_s are the weights of context relevance and semantic relevance.[6] Note that we have no requirements on how to build the affinity graphs and other intrarelation measurements can also be explored.

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

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:

$$\sum_{m=1}^{|\mathbb{I}|}\sum_{n=1}^{|\mathbb{I}|} W^I_{m,n}||\boldsymbol{i_m} - \boldsymbol{i_n}||^2 \tag{2.20}$$

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. (2.13), we obtain the overall objective function:

$$\begin{aligned}\min_{U,I,T,\mathscr{C}} g = f(\mathbf{p}^\top) \times \mathbf{1}_N + \beta(||U||^2 + ||I||^2 + ||T||^2) \\ + \alpha(tr(U^\top L_U U) + tr(I^\top L_I I) + tr(T^\top L_T T))\end{aligned} \tag{2.21}$$

where $||U||^2 + ||I||^2 + ||T||^2$ is $l-1$ regularization term to penalize large parameters, α and β are weights controlling the strength of corresponding constraints.

2.3.3 Optimization and Parameter Learning Algorithms

Next we present an algorithm to solve the optimization problem. Obviously, directly optimizing Eq. (2.21) 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 $\mathscr{C}$, respectively.*

We propose an alternating learning algorithm (ALA) to learn the factors by iteratively optimizing each subproblems. According to Theorem 1, each subproblem has a unique solution. In practice, as g is convex w.r.t. I, it is also convex w.r.t. each $\boldsymbol{i}_m$.[7] Therefore, when performing optimization on I, we optimize one row $\boldsymbol{i}_m$ at a time with other rows $\{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{m-1}, \boldsymbol{i}_{m+1}, \ldots, \boldsymbol{i}_{r_I}\}$ 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 [18] and is omitted here. We provide the proof of Theorem 2 in *Appendix* A. With the

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

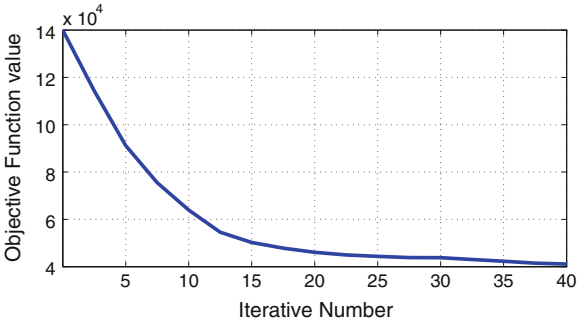


Fig. 2.3 The convergence curve of the learning algorithm. ©[2012] IEEE. Reprinted, with permission, from Ref. [34]

learnt factors, tag refinement is performed by computing the cross-space *image-tag* associations.

In the experiments, we observed that the proposed ALA converges to the minimum after about 20 iterations. Figure 2.3 shows the change of objective function values in the convergence process. We perform our experiments on MATLAB in a PC with 2.13 GHz 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 [46]. *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 nonconvex learning problems, the initialization of the factor matrices is very important to our learning algorithm. We will be working toward investigating a proper initialization scheme in the future.

2.4 Performance Evaluation

2.4.1 Dataset

We perform the experiments of social tag refinement on the large-scale web image dataset, NUS-WIDE [5]. 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.[8] The collected images belong to 50,120 unique users, with each user owning about 5 images. We select the users

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

Table 2.2 The statistics of NUS-WIDE-USER15

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

owning no less than 15 images and keep their images to obtain our experimental dataset, which is referred as NUS-WIDE-USER15. Table 2.2 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:

$$\text{Fscore} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.22}$$

2.4.2 Parameter Settings

The proposed approach, RMTF, has five parameters, the rank of factor matrices r_U, r_I, r_T and the regularization weights α, β. 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 fewer 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 [1]. The tensor $\mathscr{Y}$ 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. 2.4a

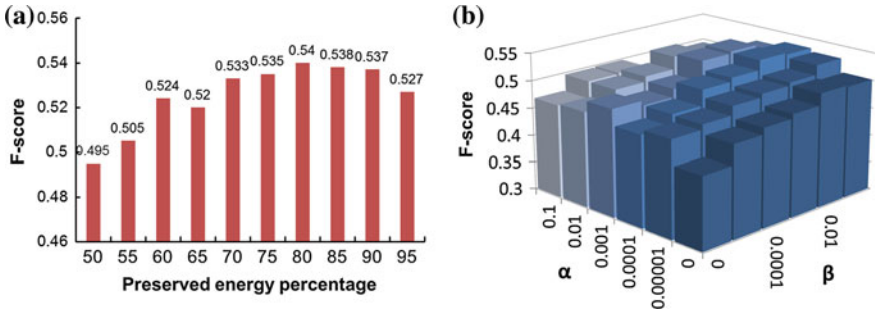


Fig. 2.4 Impact of parameters (a) rank numbers (b) α and β. ©[2012] IEEE. Reprinted, with permission, from Ref. [34]

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 α and β control how much the tensor decomposition incorporates the information of affinity intrarelations. We keep $r_U = 25$, $r_I = 105$, and $r_T = 18$. Figure 2.4b shows the impacts of α and β 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 $\beta = 0.001$.

2.4.3 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 [39].
- Tag refinement based on visual and semantic consistency (TRVSC, [20]).
- Multiedge graph (M-E Graph): a unified multiedge graph framework for tag processing proposed in [23].
- Low-Rank approximation (LR): tag refinement based on low-rank approximation with content-tag prior and error sparsity [47].
- 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 [19].

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 multicorrelation 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. (2.14) (TF_rank), and (4) TF with multicorrelation smoothness constraints, optimization under the *ranking scheme* with negative set constructed as Eq. (2.10) (RMTF).

Table 2.3 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

Table 2.3 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**

model does not fully explore the image–image intrarelations. 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 TRVSC and M-E Graph are taken from [23], 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 [31, 32] and was implemented to perform performance comparison with the proposed RMTF method. Consistent with the discussion in Sect. 2.3.2 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. 2.5. We 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," and "protest," existing methods focusing on utilizing image appearances and tag semantics fail and our

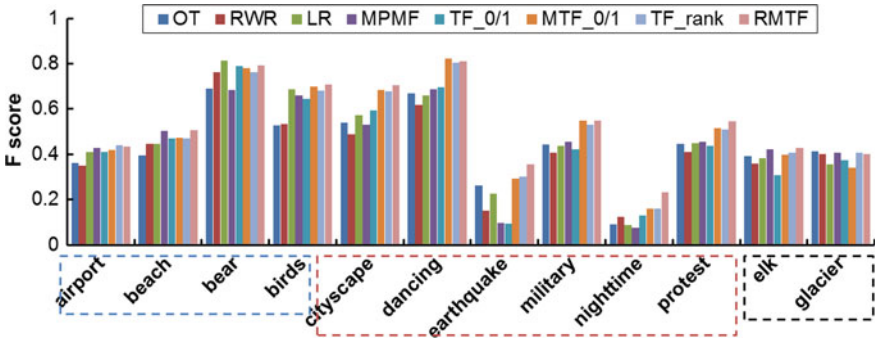


Fig. 2.5 F-score of a subset of the 81 tags for different algorithms. ©[2012] IEEE. Reprinted, with permission, from Ref. [34]

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.

2.4.4 Case Studies

We show some case studies in this subsection to demonstrate the effectiveness of RMTF. Figure 2.6 further illustrates the tag refinement results for some exemplary images by the proposed RMTF framework. For examples of Fig. 2.6c, 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. 2.6c also tagged "mosaic" and "building" to images and the tagger of Fig. 2.6d is a "sculpture" fan. Therefore, the exploited semantic is propagated into the refined results. In the original tag set of Fig. 2.6a, 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. Figure 2.6d, f further show this advantage. Moreover, Fig. 2.6b 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 contribute to a better modeling of the tagging data and derives compact *image* and *tag* factor representations.

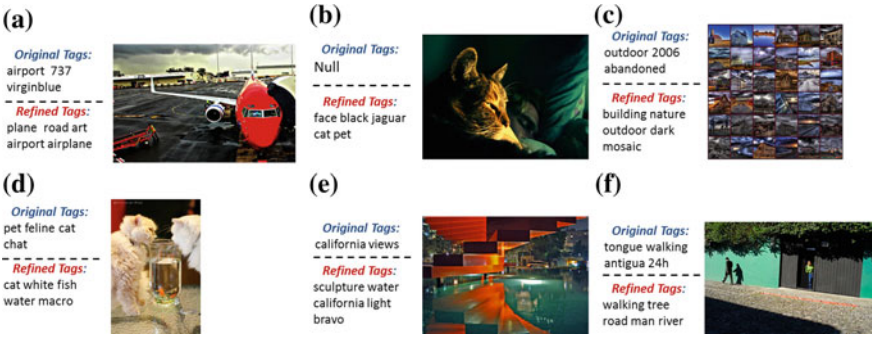
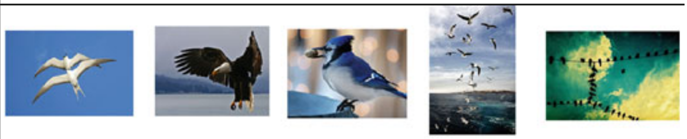


Fig. 2.6 Example of tag refinement results. For each image, the top five annotations are shown. ©[2012] IEEE. Reprinted, with permission, from Ref. [34]

Table 2.4 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 2.5 Five nearest images in the learned image subspace for each of the four selected images

Image	**Five Nearest Images**
	
	
	
	

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 Tables 2.4 and 2.5 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.

2.5 Discussions

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.

In personalized image search, the returned image results depend on not only their relevances with the query keywords, but also 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. (2.5), which reflect the users' preferences and can be leveraged to rerank the returned images.

Another potential application is personalized tag recommendation, whose goal is to predict tags for each user on a given web item (image, music, URL or publication). The reconstructed tensor $\hat{\mathscr{Y}}$ 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.

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 [27], thus the latent user interests can be understood by estimating the *user–tag* association. Actually, we have employed the derived *user* and *tag* factor matrices to build user-specific topic spaces for user modeling, and view the personalized image search problem as an reexamination into *user-tag-image* ternary correlations. Please refer the details to our recent work in [35]. Besides exploring the interrelations, we can directly evaluate the intrarelations 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. 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 book, how to model other metadata for a overall understanding is also our future work.

References

1. Acar, E., Yener, B.: Unsupervised multiway data analysis: a literature survey. IEEE Trans. Knowl. Data Eng. **21**(1), 6–20 (2009)
2. Bengio, Y., Paiement, J.-F., Vincent, P., Delalleau, O., Roux, N.L., Ouimet, M.: Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In: NIPS (2003)
3. Borghol, Y., Ardon, S., Carlsson, N., Eager, D., Mahanti, A.: The untold story of the clones: content-agnostic factors that impact youtube video popularity. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'12, pp. 1186–1194 (2012)

4. Chen, L., Xu, D., Tsang, I.W.-H., Luo, J.: Tag-based web photo retrieval improved by batch mode re-tagging. In: CVPR, pp. 3440–3446 (2010)
5. Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: CIVR (2009)
6. Cranshaw, J., Schwartz, R., Hong, J.I., Sadeh, N.M.: The livehoods project: utilizing social media to understand the dynamics of a city. In: ICWSM (2012)
7. De Choudhury, M., Sundaram, H., John, A., Seligmann, D.D.: What makes conversations interesting? Themes, participants and consequences of conversations in online social media. In: Proceedings of the 18th International Conference on World Wide Web, WWW'09, pp. 331–340 (2009)
8. Eickhoff, C., Li, W., de Vries, A.P.: Exploiting user comments for audio-visual content indexing and retrieval. In: 34th European Conference on Information Retrieval (ECIR) (2013)
9. Fang, Q., Sang, J., Xu, C., Rui, Y.: Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning. IEEE Trans. Multimed. **16**(3), 796–812 (2014)
10. Feng, W., Wang, J.: Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In: KDD, pp. 1276–1284 (2012)
11. Filippova, K., Hall, K.B.: Improved video categorization from text metadata and user comments. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'11, pp. 835–842 (2011)
12. He, X., Kan, M.-Y., Xie, P., Chen, X.: Comment-based multi-view clustering of web 2.0 items. In: Proceedings of the 23rd International Conference on World Wide Web, WWW'14, pp. 771–782 (2014)
13. Helic, D.,Strohmaier, M.: Building directories for social tagging systems. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM'10, pp. 525–534 (2011)
14. Hu, X., Tang, L., Tang, J., Liu, H.: Exploiting social relations for sentiment analysis in microblogging. In: WSDM, pp. 537–546 (2013)
15. Jin, X., Wang, C., Luo, J., Yu, X., Han, J.: Likeminer: a system for mining the power of 'like' in social media networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'11, pp. 753–756 (2011)
16. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & wordnet. In: ACM Multimedia, pp. 706–715 (2005)
17. Lappas, T., Punera, K., Sarlos, T.: Mining tags using social endorsement networks. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'11, pp. 195–204 (2011)
18. Li, W.-J., Yeung, D.-Y.: Relation regularized matrix factorization. In: IJCAI, pp. 1126–1131 (2009)
19. Li, Z., Liu, J., Zhu, X., Liu, T., Lu, H.: Image annotation using multi-correlation probabilistic matrix factorization. In: ACM Multimedia, pp. 1187–1190 (2010)
20. Liu, D., Hua, X.-S., Wang, M., Zhang, H.-J.: Image retagging. In: ACM Multimedia, pp. 491–500 (2010)
21. Liu, D., Hua, X.-S., Yang, L., Wang, M., Zhang, H.-J.: Tag ranking. In: WWW, pp. 351–360 (2009)
22. Liu, D., Hua, X.-S., Zhang, H.-J.: Content-based tag processing for internet social images. Multimed. Tool. Appl. **51**, 723–738 (2011)
23. Liu, D., Yan, S., Rui, Y., Zhang, H.-J.: Unified tag analysis with multi-edge graph. In: ACM Multimedia, pp. 25–34 (2010)
24. Liu, J., Wang, B., Li, M., Li, Z., Ma, W.-Y., Lu, H., Ma, S.: Dual cross-media relevance model for image annotation. In: ACM Multimedia, pp. 605–614 (2007)
25. Liu, X., Yan, S., Cheng, B., Tang, J., Chua, T.-S., Jin, H.: Label-to-region with continuity-biased bi-layer sparsity priors. ACM Trans. Multimed. Comput. Commun. Appl. (TOMCCAP) **8**(4), 50 (2012)
26. Lu, C., Hu, X., Chen, X., Park, J.-R., He, T., Li, Z.: The topic-perspective model for social tagging systems. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 683–692 (2010)

27. man Au Yeung, C., Gibbins, N., Shadbolt, N.: A study of user profile generation from folksonomies. In: SWKM (2008)
28. Pinto, H., Almeida, J.M., Gonçalves, M.A.: Using early view patterns to predict the popularity of youtube videos. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM'13, pp. 365–374 (2013)
29. Plangprasopchok, A., Lerman, K., Getoor, L.: Growing a tree in the forest: Constructing folksonomies by integrating structured metadata. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'10, pp. 949–958 (2010)
30. Potthast, M., Stein, B., Becker, S.: Towards comment-based cross-media retrieval. In: Proceedings of the 19th International Conference on World Wide Web, WWW'10, pp. 1169–1170 (2010)
31. Rendle, S., Marinho, L.B., Nanopoulos, A., Schmidt-Thieme, L.: Learning optimal ranking with tensor factorization for tag recommendation. In: KDD, pp. 727–736 (2009)
32. Rendle, S., Schmidt-Thieme, L.: Pairwise interaction tensor factorization for personalized tag recommendation. In: WSDM, pp. 81–90 (2010)
33. Sang, J., Liu, J., Xu, C.: Exploiting user information for image tag refinement. In: ACM Multimedia, pp. 1129–1132 (2011)
34. Sang, J., Xu, C., Liu, J.: User-aware image tag refinement via ternary semantic analysis. IEEE Trans. Multimed. **14**(3–2), 883–895 (2012)
35. Sang, J., Xu, C., Lu, D.: Learn to personalized image search from the photo sharing websites. IEEE Trans. Multimed. **14**(4), 963–974 (2012)
36. Siersdorfer, S., Chelaru, S., Nejdl, W., San Pedro, J.: How useful are your comments? Analyzing and predicting youtube comments and comment ratings. In: Proceedings of the 19th International Conference on World Wide Web, WWW'10, pp. 891–900 (2010)
37. Trevisiol, M., Jégou, H., Delhumeau, J., Gravier, G.: Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach. In: ICMR, pp. 1–8 (2013)
38. von Ahn, L., Dabbish, L.: Esp: Labeling images with a computer game. In: AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors, pp. 91–98 (2005)
39. Wang, C., Jing, F., Zhang, L., Zhang, H.: Image annotation refinement using random walk with restarts. In: ACM Multimedia, pp. 647–650 (2006)
40. Wang, C., Jing, F., Zhang, L., Zhang, H.-J.: Content-based image annotation refinement. In: CVPR (2007)
41. Xie, L., Natsev, A., Hill, M.L., Smith, J.R., Phillips, A.: The accuracy and value of machine-generated image tags: design and user evaluation of an end-to-end image tagging system. In: CIVR, pp. 58–65 (2010)
42. Xu, H., Wang, J., Hua, X.-S., Li, S.: Tag refinement by regularized lda. In: ACM Multimedia, pp. 573–576 (2009)
43. Yamamoto, T., Nakamura, S.: Leveraging viewer comments for mood classification of music video clips. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'13, pp. 797–800 (2013)
44. Ye, M., Shou, D., Lee, W.-C., Yin, P., Janowicz, K.: On the semantic annotation of places in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'11, pp. 520–528 (2011)
45. Yu, B., Ma, W.-Y., Nahrstedt, K., Zhang, H.-J.: Video summarization based on user log enhanced link analysis. In: Proceedings of the Eleventh ACM International Conference on Multimedia, MULTIMEDIA'03, pp. 382–391 (2003)
46. Zhou, Y., Wilkinson, D.M., Schreiber, R., Pan, R.: Large-scale parallel collaborative filtering for the netflix prize. In: AAIM, pp. 337–348 (2008)
47. Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: ACM Multimedia, pp. 461–470 (2010)

http://www.springer.com/978-3-662-44670-6

User-centric Social Multimedia Computing

Sang, J.

2014, XV, 108 p. 90 illus. in color., Hardcover
ISBN: 978-3-662-44670-6