Multimed Tools Appl (2018) 77:3659–3676
DOI 10.1007/s11042-017-5184-x

Social image tag enrichment based on textual similarity modeling

Miao Shen[1]

Received: 11 March 2017 / Revised: 24 August 2017 / Accepted: 30 August 2017 /
Published online: 8 November 2017

Abstract In social image sharing websites, users provide several descriptive tags to annotate their shared images. Usually, the user annotated tags are noisy, biased and incomplete. How to improve tag quality is very important for tag based applications. The content relevant tags have certain similarities or connections with each other. Thus from some highly relevant tags, we can infer the other content relevant tags for an image. In this paper, a social image tag enrichment approach is proposed. Considering the diversity of content relevant tags for the image, we first determine some seed tags which are highly relevant to image content and cover wide range of semantics. Then the seed tags are utilized to adopt semantic similarity tags for the input image. Experiments demonstrate the effectiveness of the proposed approach.

Keywords Tag enrichment · Tag ranking · Flickr · Image Annotation · Social image · Social Media

1 Introduction

There is an explosive growth in the amount of available images in social image sharing networks, such as Flickr. Social media websites allow users to provide several descriptive words called tags to illustrate the content of each shared photo [29]. The user annotated tags make the social image sharing websites better accessible to the public [1]. The social image has multimodality information [65, 70, 74], such as the tags, geo-tags, and so on. These information has close correlation with image taken. For example, the geo-tags sometimes are highly relevant to the geo-graphical information that captured by remote sensing satellites [4, 17, 28, 34–37, 39, 40, 72]. Many applications can benefit from the tags of social images [22, 35, 36, 38, 46, 52, 68, 74], for example, the pattern recognition and computer visual based applications [2, 5–8, 50, 76], personalized

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11042-017-5184-x) contains supplementary material, which is available to authorized users.

✉ Miao Shen
28855546@qq.com

[1] Xi'an Jiaotong University City College, Xi'an 710049, China

recommendations [15, 16, 19, 47, 48, 77–79], and big data based analysis. In the tag based image retrieval, user-labeled tags are noisy, biased and incomplete. The performances of tag based applications are inevitably influenced by the tag quality [27, 38].

Especially, in text based image search, if the tag is not in the tag list, it is impossible for the images ranked in to top ranked image list [38]. Thus, image tagging or tag enrichment according both the generated tags and visual information is important. Recently, many efforts have been done on image tagging to recommend more relevant tags for an image. From which more relevant tags can be inferred by textual concurrences or learning based approach.

In the proposed tag enrichment, firstly we determine the relevance of each tag to image. Then we select seed tags for this image by taking both their relevance and semantic coverage into account. Finally, tags are selected iteratively by maximizing the compatible values among tags.

The rest of this paper is organized as follows: In Section II related work on image tagging is reviewed. In Section III, the proposed approach is illustrated in detail. In Section IV experimental results and discussions are given. Conclusions are drawn in Section V.

2 Related Work

Social image tag enrichment is to enrich content relevant textual descriptions for an image from its existing tags provided by social network users based on the visual or the textual description. Social image tagging/annotation is aiming at transforming content related tags for the image. In social image, there is much multimodality information available, such as geo-locations, image taking time, user annotated tags, views, and textual comments by other social users. In social image, the tags labeled by user are noisy, biased and incomplete. From feature point of view [3, 4, 71, 79], social image tagging approaches can be classified into only visual feature, only textual feature, and their combinations.

Some tags are associated with color, texture pattern, visual content, while the other tags are from textual or local descriptor. How to select effective features for image tagging is also interesting. Recently, this problem is studied by exploring the sparsity of semantics and the low-level features [20, 24, 34, 38, 67, 70, 75]. Especially with the success of deep feature extraction approaches, such as convolution neural networks in feature, the traditional semantic gap can be further narrowed. For example, Ma et al. propose a collaborative feature selection based subspace sparsity representation approach for image annotation [12].

Model-based and model-free approaches can be adopted to fuse multimodal features in tagging. The model based approaches need to build models for each tag [9, 10, 20, 21, 24, 34, 39, 52, 65, 70, 80]. The model free approaches predict relevant tags for an image by utilizing statistical properties of tags, the low-level visual features, and other type of information [11, 46, 55, 71, 74].

In the model based approach, both discriminative and generative models can be adopted to build the relationship between tag and low-level visual features [10, 12, 20, 24, 34, 39, 65, 70, 77]. The SVM, GMM, LDA, and Multi-kernel learning based approaches are often utilized [72]. For example, Chen *et al.* use support vector machines to carry out image tagging [45]. Wang *et al.* utilize adaptive Gaussian mixture model to build visual tag dictionary [21]. Xu *et al.* refine tagging quality by utilizing regularized latent Dirichlet allocation to model the tag similarity and tag relevance [64]. Different from the model based approach, Wu *et al.* only model the hardest examples to improve tag enrichment performances, rather than building model for each tag [63].

However, this type of approaches usually meet two problems, for the number of tag is very large, it is computational intensive and complex to train model for each tag that appeared in social network. So, Gu et al. proposed to tag the community of the tags before assigning tags for each image [13].

Graph based approaches are also often adopted [11, 16, 19, 21, 23, 27, 31, 33, 43, 54, 62–64, 67, 72, 80]. A tag corresponds to a node and there is an edge between two nodes. The weight of the edge between two nodes (tags) is their similarity. The weight can be measured by taking the concurrence and visual similarity into account. Random work models are often utilized to find more relevant tags [1, 27, 29, 37, 53]. The random walk model promotes the tags with many visual similar neighbors and weakens the tags with fewer neighbors. Jia *et al.* fuse the textual similarities of tags and visual similarities of images in multi-graph reinforcement framework to find relevant tags [1]. Liu et al. model the image retagging process as a multiple graph-based multi-label learning problem [33]. They propagate the information of each tag along a tag-specific graph and determine the tag-specific visual sub-vocabulary from a collection of social images [33]. Yang et al. propose a graph model based image annotation approach by visual features to exploit the unlabeled images [67]. A multi-label classifier is trained by integrating structure learning and graph based transductive classification for providing image content relevant labels during image annotation. Some tagging approaches convert the graph based image tagging to a graph based optimization problem [42, 43, 66, 73]. Graph-cut and graph reinforcement can be utilized in finding appropriate tags for an image or video.

Except from using model based approach, some tagging approaches make full use of cooperative filtering to generate more content relevant tags. Firstly, visual near duplications for the input image are selected from large scale image corpus, and then the frequently appeared tags in the found near duplicate images are recommended to the input image [19, 24, 54, 56, 71, 74]. These approaches are based on the fact that different users label visually similar images using the same tags. However, this approach requires large scale tagged dataset. For a small dataset with nearly no duplicate it performances is inferior. Especially, for novel images which are first shared in the internet. Moreover, it only use visual features but ignores the user labeled tags.

Image tagging by mining knowledge from web is also studied in [31]. Liu *et al.* first use knowledge based method to find visual content related tags [31], then constrain the tagging vocabulary within the image content related tags. Image tagging using the supplement information of the images, such as image taking time and location, shared by social users' are also proposed recently [3, 15, 17, 18, 49, 79]. These approaches usually consist of image geo-location estimation [3], and geo-location based tagging.

3 Our Approach

In the proposed image tagging approach we classify each tag into two different sets: highly relevant to image and less relevant to image. Our goal is to recommend tags iteratively. Firstly some seed tags which are highly relevant to image content are selected based on which we extract more content relevant tags. Then the top recommended tags in previous iteration are served as seeds to find new relevant tags iteratively.

There are two major problems needed to be solved in the proposed tag enrichment approach which are summarized as follows:

1) How to select seed tags, i.e. how to determine the relevance of tag to image.
2) How to recommend tags covering diverse semantics based on the top relevant tags to widen the semantic coverage the image.

3.1 Tag Enrichment

The sketch map of the proposed tag enrichment approach is illustrated in Fig. 1. S_r denotes the determined most relevant tag set with the tag ranks smaller than or equals to r, and T_r denotes the tag set with the ranks of the tags larger than r. the black dark shapes denote the seed tags, the light gray shapes denote the tags to be adopted and the green shapes denote the adopted tags. In the proposed tag enrichment approach, a new tag x will be extracted from $T_{r\text{-}1}$will be assigned with rank r based on the $S_{r\text{-}1}$.

3.1.1 Similarity of Tag to Image

The tag and image are with two different modalities. It is difficult to measure their similarity directly. However, taking the advantages of the social image with surrounding tags, in this paper the similarity of the tag τ to the image I (denoted by $R(\tau)$) is measured by the similarity of image I to all the images containing the tag τ as follows:

$$R(\tau) = \frac{1}{|\Theta_\tau|} \sum_{x\in\Theta_\tau} \exp\left(-\|F_I - F_x\|^2 \Big/ \sigma^2\right) \tag{1}$$

where Θ_τ denotes the image set that contains tag τ, the image number in this set is $|\Theta_\tau|$. σ^2 is set to be the median value of all the pair wise Euclidean distances between images [29]. F_I and F_x are the visual features of images I and x. It can be both low-level visual features or deep features extracted from CNN [76].

3.1.2 Similarity of Tag to Tag

Tag to tag distance is measured by the commonly utilized Google distance [38, 66] which is expressed as follows:

$$d(p,q) = \frac{\max(\log f(p), \log f(q)) - \log f(p,q)}{\log W - \min(\log f(p), \log f(q))} \tag{2}$$

where $d(p,q)$ is the normalized Google distance of tag p and q, $f(p)$ and $f(q)$ are the numbers of images containing tag p and q. $f(p, q)$ is the number of images containing both the tags p and q. In this paper, these numbers are obtained by performing search by tag on Flickr website using

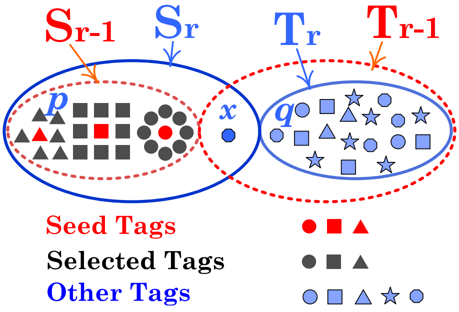

Fig. 1 Sketch map of tag enrichment. S_r denotes the tag set with the tag ranks smaller than or equals to r, and T_r denotes the tag set with the ranks of the tags larger than r. Based on the $S_{r\text{-}1}$, a new tag x selected from $T_{r\text{-}1}$will be assigned with rank r

the tags as query terms [29]. W is the total number of images on Flickr. Correspondingly, the tag to tag score is expressed as follows:

$$s(p,q) = \exp(-d(p,q)) \tag{3}$$

3.1.3 Problem Descriptions

Let $\Gamma = \{t_1, \cdots, t_{|\Gamma|}\}$ denote a tag set, with its tag number denoted by $|\Gamma|$. Our goal is to rank the tag with high relevance ahead of the tags with low relevance. The smaller the rank means the higher the relevance of the tag to the image.

Assuming that the most relevant (r-1) tag list $S_{r\text{-}1}$ has already been determined (**the detailed approach can be found in the Algorithm 1**), our tag enrichment algorithm aims at finding a candidate tag x ($x \in T_{r-1}$) to be with rank r ($r < |\Gamma|$). From the sketch map as shown in Fig. 1, the relationship of the sets S_r, T_r, $S_{r\text{-}1}$, and $T_{r\text{-}1}$ is as follows

$$\begin{cases} S_{r-1} + T_{r-1} = \Gamma, & S_r + T_r = \Gamma \\ S_{r-1} + x = S_r, & T_{r-1} - x = T_r \end{cases} \tag{4}$$

The corresponding tag numbers in S_r and T_r have following relationship

$$\begin{cases} |S_r| = r, & |T_r| + |S_r| = |\Gamma| \\ |T_r| - |T_{r-1}| = -1, & |S_r| - |S_{r-1}| = 1 \end{cases}, r \in \{1, \cdots, |\Gamma|\} \tag{5}$$

It is an optimizing problem to determine which tag in T_r can be selected as the r-th relevant tag to the image. Assigning the tag x with rank r should take the cost of transferring it from $T_{r\text{-}1}$ to S_r and the cost of assigning the tag x be the r-th relevant tag to the image. Let $E^{(r)}(L)$ denote the compatible value (or dissolvability) of assigning the tag x with rank r on the base of the already determined (r-1) relevant tags in $S_{r\text{-}1}$ and the other tags in $T_{r\text{-}1.}$ Thus $E^{(r)}(L)$ consists of two terms as follows

$$E^{(r)}(L) = E_n^{(r)}(L) + E_I^{(r)}(L) \tag{6}$$

where $L = \{L_p | p \in \Gamma\}$ is an iteration related labeling of tag setΓ. $L_p = s$(or t) means the tag p is relevant (or irrelevant) to the image at step r. In Eq.(6), $E_n^{(r)}(L)$ is the overall compatible value of partitioning the tags into two sets.

$$E_n^{(r)}(L) = \sum_{\{p,q\} \in \Gamma} V_{p,q}^{(r)}(L_p, L_q) \tag{7}$$

$V_{p,q}^{(r)}(L_p, L_q)$be the compatible value of assigning tag p and q with labels L_p and L_q at r-th iteration. $V_{p,q}^{(r)}(L_p, L_q)$ can be measured by the textual similarity in terms of the often utilized normalized Google distances.

The larger the Google distance of tag p and tag q, the smaller the compatible value. According to the normalized Google distance, the compatible value of tag p and tag q by assigning them labels L_p and L_q can be measured as follows

$$V_{p,q}(L_p, L_q) = \begin{cases} \exp(-d(p,q)) & \textit{if } L_p = L_q \\ 1 - \exp(-d(p,q)) & \textit{otherwise} \end{cases} \tag{8}$$

Especially, if $p = q$, then the compatible value $V_{p,p}(L_p, L_p) = 1$. Moreover, we have $V_{p,q}(L_p, L_q) = V_{q,p}(L_q, L_p)$.

In Eq.(6) $E_I^{(r)}(L)$ is the overall compatible value of assigning the tags with label $\boldsymbol{L}$ at step r for the input image. It can be expressed by summing up the compatible values of each tag to image. Thus it can be written as follows:

$$E_I^{(r)}(L) = \sum_{p \in \Gamma} D\left(L_p\right) \tag{9}$$

Where $D(L_p)$ be the compatible value of assigning the tag p with label L_p.

$$D\left(L_p\right) = \begin{cases} R(p), & if\ L_p = s \\ 1-R(p), & if\ L_p = t \end{cases} \tag{10}$$

where $R(p)$ measures the similarity of the tag p to the image with respect to Eq.(1), $s = 0$ and $t = 0$.

3.1.4 Iterative Optimal Tag Selection

The optimal tag can be determined by maximizing the variation of the compatible values of adopting a tag x between two adjacent steps r-1 and r. We determine the optimal solution by maximizing the variations of compatible values of changing the label of tag x from irrelevant to relevant at step r. The compatible value variations $\Delta E(x)$ of changing the label of tag x from irrelevant to relevant at any two neighboring steps r and r-1 is calculated as follows

$$\begin{aligned} \Delta E(x) &= E^{(r)}(L) - E^{(r-1)}(L) \\ &= \left(E_n^{(r)}(L) + E_I^{(r)}(L)\right) - \left(E_n^{(r-1)}(L) + E_I^{(r-1)}(L)\right) \\ &= \left(E_n^{(r)}(L) - E_n^{(r-1)}(L)\right) + \left(E_I^{(r)}(L) - E_I^{(r-1)}(L)\right) \\ &= \Delta E_n^{(r)}(x) + \Delta E_I^{(r)}(x) \end{aligned} \tag{11}$$

Eq.(11) can be rewritten as follows:

$$\Delta E(x) = \Delta D(x) + 2\Delta V_1(x) + 2\Delta V_2(x) \tag{12}$$

where $\Delta D(x)$ is the compatible value variation when changing the label of the tag x from t to s (i.e. from irrelevant to relevant, or changing its label from 0 to 1) at the r-th step.

$$\Delta D(x) = D(L_x = s) - D(L_x = t) = 2R(x) - 1$$

$\Delta V_1(x)$is the variation of compatible value of the tags in $S_{r\text{-}1}$ to tag x when the label of tag x is changed from t to s at step r.

$$\begin{aligned} \Delta V_1(x) &= \sum_{p \in S_{r-1}, q \in \Gamma} V_{p,q}^{(r)}\left(L_p = s, L_q\right) - \sum_{p \in S_{r-1}, q \in \Gamma} V_{p,q}^{(r-1)}\left(L_p = s, L_q\right)\ \Delta V_1(x) \\ &= \sum_{p \in S_{r-1}} V_{p,x}\left(L_p = s, L_x = s\right) - \sum_{p \in S_{r-1}} V_{p,x}\left(L_p = s, L_x = t\right) \end{aligned}$$

$\Delta V_2(x)$is the variation of compatible value of the tags in T_r to tag x when the label of tag x is changed from t to s.

$$\begin{aligned} \Delta V_2(x) &= \sum_{p \in T_r, q \in \Gamma} V_{p,q}^{(r)}\left(L_p = t, L_q\right) - \sum_{p \in T_r, q \in \Gamma} V_{p,q}^{(r-1)}\left(L_p = t, L_q\right) \\ &= \sum_{p \in T_r} V_{p,x}\left(L_p = t, L_x = s\right) - \sum_{p \in T_r} V_{p,x}\left(L_p = t, L_x = t\right) \end{aligned}$$

Assigning a tag x_0 with rank r can bring maximum compatible value improvement. Thus, We have:

$$\begin{aligned} x_0 &= \arg\max_{x\in T_{r-1}} \Delta E(x) = \arg\max_{x\in T_{r-1}} \left(\Delta E_n^{(r)}(x) + \Delta E_I^{(r)}(x)\right) \\ &= \arg\max_{x\in T_{r-1}} \left(\Delta D(x) + 2\times \Delta V_1(x) + 2\times \Delta V_2(x)\right) \end{aligned}, \; x_0\in T_{r-1}, \quad x_0 = S_r\cap T_{r-1} \quad (13)$$

That is to say that changing x_0 from irrelevant to relevant can maximize the compatible value. That is to say, the tags in $S_{r\text{-}1}$ like to adopt x_0 into the same set and at the same time the tags in $T_{r\text{-}1}$ most likely to kick x_0 out of their set at step r.

3.1.5 Seed Tags Selection

The multiple seed tags are selected by taking into both the relevance of tag to image and the diversities of top ranked tags. A greedy based searching approach is utilized to select seed tags by maximizing the semantic coverage of the seed tags iteratively. Assuming that the most relevant m-1 seed tags are determined and pushed in the list $S_{m\text{-}1}$ (**the detailed determination approach can be found in the Algorithm 1**), the belief value of the tag τ be selected as the m-th seed tags for the input image I should take both the relevance of the tag to the image and the semantic compensations of the tag τ to the already determined m-1 seed tags. The belief value of the tag τ is represented as follows:

$$B(\tau) = R(\tau)\times C(\tau) \quad (14)$$

where $C(\tau)$ is the semantic compensations of tag τ to the tag set Γ. In this paper, we use the minimum score of tag τ to the tags in Γ to measure the semantic compensations $C(\tau)$ as follows.

$$C(\tau) = \min_{\tau_i\in\Gamma}\left(1 - s(\tau,\tau_i)\right) \quad (15)$$

where $s(\tau,\tau_i)$ is the textual similarity score of the tag τ and the tag τ_i which is given as follows.

$$s(\tau,\tau_i) = \exp(-d(\tau,\tau_i)) \quad (16)$$

where $d(\tau,\tau_i)$ is the normalized Google distance of tag τ and τ_i, which is expressed in Eq.(2). The corresponding seed tags selection algorithm is as shown in Algorithm 1:

3.2 Flowchart of Image Tagging

Given image I with the initial tag set $\varphi = \{\tau_1, \cdots, \tau_{|\varphi|}\}$, let $\Omega = \{v_1, \cdots, v_{|\Omega|}\}$ denote the set of the whole tag vocabulary, which is obtained from the crawled Flickr images. The total tag number in Ω is $|\Omega|$. Our image tagging approach consists of the following steps: 1) feature extraction and determine the relevance of tag to image; 2) select seed tags for the image as shown in Algorithm 1; 3) using textual similarity to generate new tags for the image. Now we give the flowchart of the proposed image tagging approach in Algorithm 2.

4 Experiments and Discussion

To evaluate the performances of the proposed tagging approach, we conduct experiments on a crawled dataset using selected 25 queries and NUS-Wide [9]. We compare our approach (denoted TS) with the raw tags by Flickr users (denoted INIT), tag concurrence based approach (denoted COCR), random walk based tag ranking approach [29] (denoted RANK), and visual neighbor voting (denoted NBVT) [24]. Our experiments consist of the raw tag ranking, tag enrichment and tag based image retrieval.

4.1 Dataset

In our crawled dataset, we select 25 queries, including *alcedoatthis*, *apple*, *beach*, *bear*, *butterfly*, *cherry*, *deer*, *eagle*, *forest*, *highway*, *jeep*, *lavender*, *lotus*, *orange*, *peacock*, *rose*, *sailship*, *sea*, *sky*, *strawberry*, *sun*, *sunflower*, *tiger*, *tower*, and *zebra*, then perform tag-based image search with “ranking by interestingness” option on Flickr. The top 5000 returned images for each query are collected together with their associated information, including tags, uploading time, user identifier, etc. In this way, we obtain a social image collection consisting of 52,418 images. Totally, there are 887,353 raw tags.

4.1.1 Preprocessing

We match each tag with the entries in a Wikipedia thesaurus and we keep only the tags that have a coordinate in Wikipedia [38, 21, 57, 66]. We adopt a simple and effective method by removing the tags with their frequency that appear less than 20 times. Also some high frequency stop words, are removed as irrelevant tags [15, 16].

4.1.2 Feature Representation

For each image, we extract 470-dimensional features, including 225-dimensional block-wise color moment features generated from 5-by-5 fixed partition of the image, 170-dimensional hierarchical wavelet packet features [1], and 75-dimensional edge distribution histogram features.

In this paper, we only utilize the basic low-level features in our social image tagging approaches. Actually, high level feature extraction approach that determined from the CNN can be also utilized.

4.2 Evaluation of Tag enrichment

We use the average relevant score of the test images for performance evaluation as that utilized in [44]. 200 images are randomly selected from our Flickr dataset for labeling by five persons. For each image, each of its user labeled tags is labeled as one of the five levels: Most Relevant (score 4), Relevant (score 3), Partially Relevant (score 2), Weakly Relevant (score 1), and Irrelevant (score 0).

1) the tag is with **Most Relevant** to the image if most important content of the image is disclosed by it.

2) the tag is with **Relevant** to the image if important content but not the most important content is disclosed by it.
3) the tag is with **Partial Relevant** to the image if some parts of image content is disclosed by it.
4) the tag is with **Weak Relevant** if a small part of image content is disclosed by it.
5) the tag is with **Irrelevant** to the image if no content of the image is disclosed by it.

Given an image with ranked tag list $\{t_1, \cdots, t_n\}$, the average relevant score is computed as

$$AVR_n = \frac{1}{K}\sum_{k=1}^{K} R_n^k \tag{16}$$

where R_n^k is the relevance level of the n-th tag of k-th test image and K is the total test image number.

4.2.1 Tag ranking Performance

Fig. 2 (a) shows the average relevant scores of valid tags of the user annotated raw tags on Flickr (denoted INIT), ranked results of tags by random walk [29] (denoted RANK), visual neighbor voting [24] (denoted NBVT), and the proposed approach (denoted TS) at depths in the range [29, 70]. For the first ranked tags, the average relevant scores of INIT, COCR, NBVT, RANK, RLVT and TS are 2.56, 2.76, 2.90, 2.89, 3.01 and 3.01 respectively. From relevance ranking the performances of the first tag are highly improved. As the first tag of RLVT and TS are both selected by the most relevant, their performances are the same. Under @5, the average relevant scores of them are 2.17, 2.22, 2.43, 2.43, 2.49, and 2.51 respectively. From Fig. 2, we find that the tag ranking performances of TS are better than these of NBVT, RANK and INIT. Moreover, the improve algorithm of similar compatible approach by using the diverse semantic, better performances are achieved. This is due to the fact that, the preselected tags can cover wide range of semantics.

4.2.2 Tag Enrichment Performance

Fig. 2 (b) shows the average relevant scores of enriched tags by COCR, RANK, NBVT, RLVT and TS at depths in the range [29, 70]. For the first ranked tags, the average relevant scores of COCR, NBVT, RANK, RLVT and TS are 2.72, 2.98, 2.87, 3.23, and 3.23 respectively. From relevance ranking the performances of the first tag are highly improved. Due to the fact that the first tag of RLVT and TS are both selected by the most relevant, their performances are the same. Under @5, the average relevant scores of them are 2.49, 2.64, 2.74, 2.91, and 2.99 respectively. From Fig. 2, we find that the tag ranking performances of TS are better than these of NBVT, RANK and RLVT. Moreover, the improve algorithm of similar compatible approach by using the diverse semantic, better performances are achieved. This is due to the fact that, the preselected tags can cover wide range of semantics. The subjective tag enrichment results for some social images with their initial tags are given in Table 3. From Table 3, we find that comparatively better performances are achieved by ours.

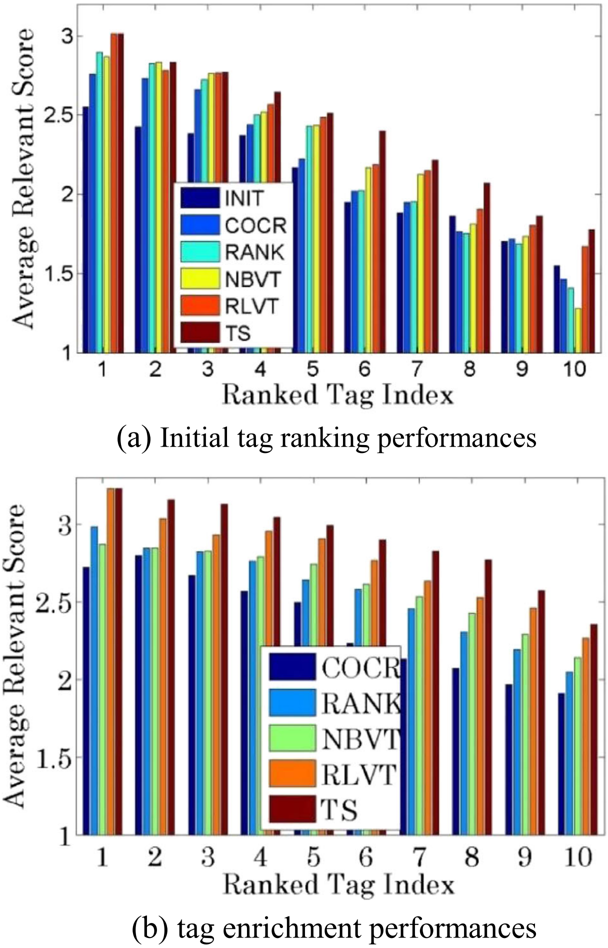

Fig. 2 top 10 image tagging performance comparisons for INIT, COCR, RANK, NBVT, RLVT and TS. **a** Initial tag ranking performances. **b** tag enrichment performances

4.3 Relative Performances

In Part C of Section III, we provide the relative tag ranking performance for the raw tags and three tag ranking approaches: RANK, NBVT and TS. Our approach can also improve tag enrichment performances if the enriched tag list of an image is provided. In this Section, we evaluate the relative tag enrichment performances of TS over RANK and NBVT.

We invite five volunteers to evaluate tag ranking performances for the compared two approaches into following three degrees: 1) **better than**, 2) **almost the same**, and 3) **poor than**. And then, we use weighted average precision (WAP) to evaluate the relative performances of two approaches as follows.

$$WAP = (b-p)\Big/(b+s+p) \qquad (17)$$

where b, s, p denote the ratio of "better than", "almost the same", and "poor than" respectively.

The relative performances of TS over RANK and NBVT at NDCG depth @1, @5 and @10 are shown in Table 1. The corresponding ratios of "better than", "almost the same", and "poor than" under the above NDCG depth are also given. As TS and RANK both select the most relevant tag as the first rank, the performances of TS and RANK @1 are identical. With the increase of NDCG depth, TS outperforms RANK 37% and 43% respectively under @5 and @10. By comparing TS with NBVT, the WAP values of TS over NBVT are 24.2%, 44.2% and 49.2% respectively.

4.4 Image Search

We conduct image search to verify the effectiveness of proposed tag filtering approach our crawled dataset. We first select the 25 queries from crawl our dataset, and we compare the tag based image search results based on the enriched tags by following methods:(1) Image Search with raw tags labeled by user (INIT).(2) Image Search with tags enriched by TS (TS).(3) Image Search with tags enriched by RANK (RANK).(4) Image Search with tags enriched by NBVT (NBVT).

To compare the image search results, we obtain the ranked image lists of different approaches for each query. We invited 5 subjects to label the relevance of the top 50 results for each query and each method. For each ranking list, the images are decided as relevant or irrelevant with respect to the query terms by the five volunteers. We use average precision (AP) as image search evaluation metric. Give a ranked image list, the AP at depth n is defined as follows

$$AP_n = \frac{1}{n} \sum_{j=1}^{n} R_j \tag{18}$$

where R_j measures the relevance of the j-th instance to the query. $R_j = 1$ if the j-th instance is relevant and 0 otherwise. To evaluate the overall performance, we use mean average precision (MAP) of the 25 queries (Fig. 3).

Figure 4 illustrates the MAP measurement at different return depths on the crawled dataset. We can see that the search results on the enriched tags by RANK, NBVT, RLVT and TS are better than the INIT. Moreover, our approach TS outperforms the RANK, NBVT and RLVT [44]. This shows that our approach can enrich image content relevant tags for each image.

Table 1 Statistical results of the relative performances of TS over RANK and NBVT for the 500 test images

	Performance	@1	@5	@10
TS Over RANK	better than	0	46.6%	49.2%
	almost the same	500	43.8%	44.6%
	Poor than	0	9.6%	6.2%
	WAP	**0.0%**	**37.0%**	**43.0%**
TS Over NBVT	better than	26.6%	52.4%	53.4%
	almost the same	71.0%	39.4%	42.4%
	Poor than	2.4%	8.2%	4.2%
	WAP	**24.2%**	**44.2%**	**49.2%**

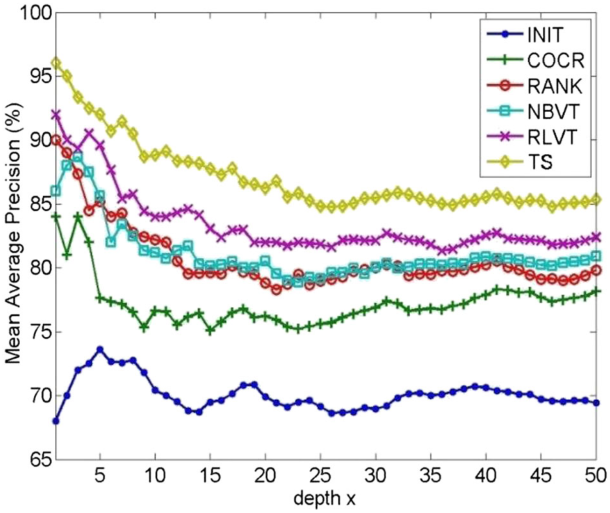

Fig. 3 Tag based image search results of different approaches under depth x in the range [29, 41]

4.5 Discussion

In [44], the authors only use the most relevant tags as the seed tags for tagging a social image. In this section, we go to discuss whether selecting seed tags can improve tagging performance. Fig. 5 shows the tagging performances of using one, two, three, four and seed number determination approach by determined by the adaptive seed tag selection (denoted ASTS) shown in Algorithm 1. In Fig. 5 the performances of using one seed tag is actually the [44]. From this figure we find that manually setting the seed number to be 2 can achieve better performances. However, the adaptive seed tag selection approach achieves the best performances.

Fig. 5 shows the performances of determining seed tags from user annotated initial tag, top 20 tags recommended by NBVT, and COCR then we use ASTS to determine the seed tags, and tags determined by ASTS. It is interesting to find that user annotated tags are not as good

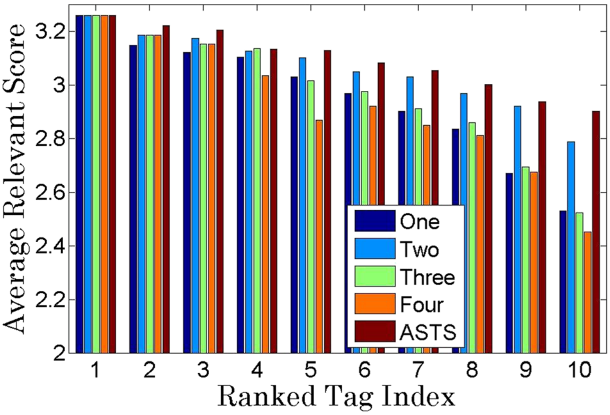

Fig. 4 Image Tagging performance of TS using one, two, three, four seed tags and adaptive seed tag selection algorithm

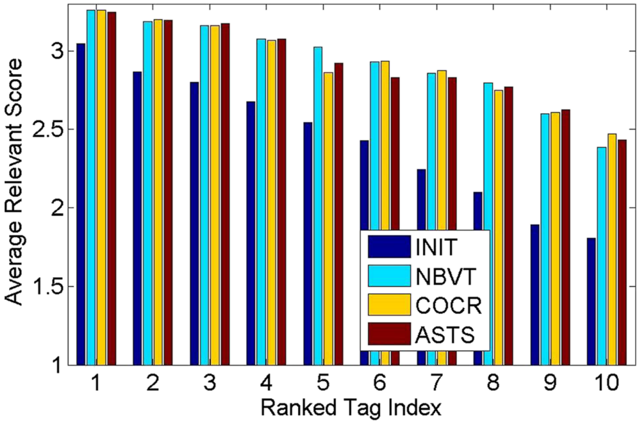

Fig. 5 Image Tagging performance of TS by selecting tags from user labeled tags, tags enriched by NBVT, COCR and ASTS

as the tags by the existing tag enrichment approach. However, different enrichment approach does not influence significantly. Table 3 shows some flickr images and the enriched tags by the corresponding different INIT, RANK, and NBVT. The comparisons show the effectiveness of the proposed approach.

For example, for the second image, the NBVT approach miss the key tag “women” and some false tag “butterfly”, while our approach can generate more content relevant tags such as, “fashion” and “art”.

4.6 Experiments on NUS-Wide

The NUS-Wide dataset is composed with two parts: the training part, which contains 27,807 images, and testing part, which contains 27,808 images [9]. All images are manually annotated with the concepts from 81 Ground Truth. Thus, in this paper, Recall, Precision and F1 are used to measure the performance of different image tagging approach. Except the images and the ground truth labels for each image, the low-level features extracted from the image including color histogram (64D), color correlation histogram (73D), edge-detection histogram (73D), block-wised color moments (256D) and wavelet texture (128D) are also provided. We use the features provided by NUS-Wide, rather than those of ourselves.

There are no initial tags for the test and training image, thus in this section we only using the visual features for tag recommendation. Correspondingly, we only use the visual features for determining the relevant tags. Thus we can only provide the performances of RANK,

Table 2 The mean of the average Recall, Precision and F1 values of the top 5 tags by RANK, NBVT, RLVT and TS on NUS-Wide dataset

	Top 5 ranked tags		
method	Recall	Precision	F1
RANK	30.10	15.79	20.71
NBVT	31.17	18.19	22.97
RLVT	40.63	19.17	26.05
TS	40.73	19.28	26.17

Table 3 Raw tags (the 2nd column) annoted by Flickr users for the input images (as shown in the first column), and the ranked initial tag lists by random walk model (denoted RANK, the 3rd column), visual neighbor voting (denoted NBVT, the 4-th row), the first 10 tags enriched by RANK (the 6-th column), and the tag lists of TS over RANK (the 7-th column)

Photo	INIT	RANK	NBVT	TS	RANK	NBVT	TS
	insect lavender bee pollen	bee pollen insect lavender	insect lavender pollen bee	bee lavender pollen insect	pollen plant garden bee insect flower green yellow bokeh color	lavender purple flower garden bee bokeh nature blue green insect	bee insect pollen garden plant flower bokeh green yellow color
	portrait woman girl beauty glamour	woman girl portrait glamour beauty	portrait beauty girl woman glamour	woman beauty glamour girl portrait	woman girl portrait beauty art fashion photography man color	bird blue nature green animal butterfly feather zoo wildlife color	girl woman portrait beauty fashion photography art color man
	sea portrait beach smile reflex spring uro	sea beach smile portrait	sea beach portrait smile	sea beach smile portrait	beach sand ocean sea water waves woman man holiday shore	sea beach ocean water blue sky nature sand nikon landscape	beach sea woman water sand ocean waves shore holiday man
	love butterfly icarus polyommatus	love butterfly	butterfly love	butterfly love	love beauty butterfly life nature dream black bokeh garden yellow	garden nature black wild butterfly yellow bokeh flower gray flowers	love butterfly beauty life garden nature yellow bokeh black dream
	tree pinetree pine oregon forest truck logging lumber pinaceae	pine forest tree oregon truck	forest tree pine oregon truck	pine forest tree oregon truck	pine forest tree oregon truck moss trail mist fog landscape	forest nature green tree landscape wood nikon water sun germany	pine forest tree oregon truck moss trail mist fog landscape
	river cross tiger	cross river tiger	tiger river cross	river cross tiger	cross river travel water nikon tiger digital asia eos portrait	tiger zoo animal cat wildlife nature nikon india tigris mammal	river cross tiger water travel sitting portrait digital eos nikon asia

RLVT, NBVT, and TS. In order to make fair comparisons, the features utilized by RANK, RLVT, NBVT and TS are all the same. All the five low-level features are utilized; the total dimension of the feature of each image is 594. Moreover the mean of average Recall, Precision and F1 values are shown in Table 2 respectively. From Table 2, we find that the performances of TS still outperform RLVT a little. This is because the concepts in NUS-Wide are relative independent. This is caused by the fact that the concepts in NUS-Wide dataset are manually labeled.

5 Conclusion

In this paper, a tag filtering approach based on textual similarity modeling is proposed. This approach classifies raw tags into two sets. The relevant tags are determined one by one by changing their labels from irrelevant to relevant. The tag ranking problem is converted to an optimization problem. The energy function takes both the compatible value of this tag to the tags from the two set, and cost of assigning this tag to be relevant to the image content. Experimental results show the effectiveness of the proposed approach (Table 3).

References

1. Ames M, Naaman M (2007) Why We Tag: Motivations for Annotation in Mobile and Online Media. In Proc. SIGCHI Conference on Human Factors in Computing System
2. Chang X, Yang Y (2016) Semi-supervised Feature Analysis by Mining Correlations among Multiple Tasks. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2016.2582746
3. Chang X, Nie F, Wang S, Yang Y, Zhou X, Zhang C (2016) Compound Rank-k Projections for Bilinear Analysis. IEEE Trans Neural Netw Learn Syst 27(7):1502–1513
4. Chang X, Nie F, Yang Y, Zhang C, Huang H (2016) Convex Sparse PCA for Unsupervised Feature Analysis. ACM Trans Knowl Discov Data 11(1):3:1–3:16
5. Chang X, Nie F, Wang S, Yang Y, Zhou X, Zhang C (2016) Compound Rank-k Projections for Bilinear Analysis. IEEE Trans Neural Netw Learn Syst 27(7):1502–1513
6. Chang X, Ma Z, Yang Y, Zeng Z (2017) Alexander G. Hauptmann: Bi-Level Semantic Representation Analysis for Multimedia Event Detection. IEEE Trans Cybern 47(5):1180–1197
7. Chang X, Yu Y, Yang Y, Xing EP (2017) Semantic Pooling for Complex Event Analysis in Untrimmed Videos. IEEE Trans Pattern Anal Mach Intell 39(8):1617–1632
8. Chang X, Ma Z, Lin M, Yang Y, Hauptmann AG (2017) Feature Interaction Augmented Sparse Learning for Fast Kinect Motion Detection. IEEE Trans Image Process 26(8):3911–3920
9. Chua T, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) Nus-wide: A real-world web image database from national university of Singapore. In Proc. CIVR
10. Datta R, Joshi D, Li J, Wang JZ (2007) Tagging over time: Realworld image annotation by lightweight meta-learning. In: Proc. ACM Mutlimedia, p 393–402
11. Feng S, Lang C, Xu D (2010) Beyond tag relevance: integrating visual attention model and multi-instance learning for tag saliency ranking. CIVR, p 288–295
12. Gao Y, Wang M, Zha Z, Shen J, Li X, Wu X (2013) Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search. IEEE Trans Image Process 22(1):363–376
13. Gu Y, Qian X, Li Q, Wang M, Hong R, Tian Q (2015) Image Annotation by Latent Community Detection and Multi-Kernel Learning. IEEE Trans Image Process 24(11):3450–3463
14. Han Y, Wu F, Tian Q, Zhuang Y (2012) Image Annotation by Input-Output Structural Grouping Sparsity. IEEE Trans Image Process 21(6):3066–3079
15. Jiang S, Qian X, Shen J, Fu Y, Mei T (2015) Author Topic Model-Based Collaborative Filtering for Personalized POI Recommendations. IEEE Trans Multimedia 17(6):907–918
16. Jiang S, Qian X, Fu Y, Mei T (2016) Personalized Travel Sequence Recommendation on Multi-Source Big Social Media. IEEE Trans Big Data 1(2):43–56
17. Joshi D, Luo J, Yu J, Lei P, Gallagher A (2011) Using Geotags to Derive Rich Tag-Clouds for Image Annotation, Social Media Modeling and Computing. Springer, Berlin
18. Kleban J, Moxley E, Xu J, Manjunath BS (2009) Global annotation on georeferenced photographs. In: Proc. CIVR
19. Lei X, Qian X, Zhao G (2016) Rating Prediction based on Social Sentiment from Textual Reviews. IEEE Trans Multimedia 18(9):1910–1921

20. Li J, Wang JZ (2008) Real-time computerized annotation of pictures. IEEE Trans Pattern Anal Mach Intell 30(6):985–1002
21. Li J, Qian X, Lan K, Qi P, Sharma A (2015) Improved image GPS location estimation by mining salient features. Sig. Proc.: Image Comm. 38:141–150
22. Li X, Chen L, Zhang L, Ma W, Lin F (2006) Image annotation by large-scale content-based image retrieval. ACM MM
23. Li X, Snoek CGM, Worring M (2008) Learning tag relevance by neighbor voting for social image retrieval. ACM MIR, p 180–187
24. Li X, Snoek C, Worring M (2009) Learning Social Tag Relevance by Neighbor Voting. IEEE Trans Multimedia 11(7):1310–1322
25. Li G, Wang M, Lu Z, Hong R, Chua T (2012) In-Video Product Annotation with Web Information Mining. ACM Trans Multimed Comput Commun Appl 8(4)
26. Li J, Qian X, Tang Y, Yang L, Mei T (2013) GPS estimation for places of interest from social users' uploaded photos. IEEE Trans Multimedia 15(8):2058–2071
27. Li X, Guo Q, Lu X (2016) Spatiotemporal Statistics for Video Quality Assessment. IEEE Trans Image Process 25(7):3329–3342
28. Li X, Mou L, Lu X (2016) Surveillance Video Synopsis via Scaling Down Objects. IEEE Trans Image Process 25(2):740–755
29. Liu D, Hua X, Yang L, Wang M, Zhang H (2009) Tag ranking. In: Proc. WWW
30. Liu D, Hua X-S, Wang M, Zhang H-J (2010) Retagging social images based on visual and semantic consistency. In: Proc. ACM WWW, p 1149–1150
31. Liu D, Hua X-S, Wang M, Zhang H-J (2010) Image retagging. In: Proc. ACM Multimedia
32. Liu D, Wang M, Hua X, Zhang H (2011) Semi-Automatic Tagging of Photo Albums via Exemplar Selection and Tag Inference. IEEE Trans Multimedia 13(1):82–91
33. Liu D, Yan S, Hua X, Zhang H (2011) Image Retagging Using Collaborative Tag Propagation. IEEE Trans Multimedia
34. Lu X, Li X (2014) Multiresolution Imaging. IEEE Trans Cybern 44(1):149–160
35. Lu X, Wang Y, Yuan Y (2013) Graph Regularized Low-Rank Representation for Destriping of Hyperspectral Images. IEEE Trans Geosci Remote Sens 51(7):4009–4018
36. Lu X, Wu H, Yuan Y, Yan P, Li X (2013) Manifold Regularized Sparse NMF for Hyperspectral Unmixing. IEEE Trans Geosci Remote Sens 51(5):2815–2826
37. Lu X, Li X, Li M (2015) Semi-Supervised Multi-task Learning for Scene Recognition. IEEE Trans Cybern 45(9):1967–1976
38. Lu D, Liu X, Qian X (2016) Tag based Image Search by Social Re-Ranking. IEEE Trans Multimedia 18(8): 1628–1639
39. Lu X, Yuan Y, Zhang X (2016) Jointly Dictionary Learning for Change Detection in Multispectral Imagery. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2016.2531179
40. Lu X, Li X, Zheng X (2017) Latent Semantic Minimal Hashing for Image Retrieval. IEEE Trans Image Process 26(1):355–368
41. Mei T, Wang Y, Hua X, Gong S, Li S (2008) Coherent image annotation by learning semantic distance. In: Proc. CVPR
42. Moxley E, Mei T, Manjunath B (2010) Video annotation through search and graph reinforcement mining. IEEE Trans Multimedia 12(3):184–193
43. Qian X, Hua X (2011) Graph-cut based tag enrichment. In: Proc. SIGIR, p 1111–1112
44. Qian X, Hua X, Hou X (2012) Tag Filtering based on Similar Compatible Principle. In: Proc. ICIP
45. Qian X, Liu X, Zheng C, Du Y, Hou X (2013) Tagging photos using users' vocabularies. Neurocomputing 111:144–153
46. Qian X, Hua X, Tang Y, Mei T (2014) Social Image Tagging with Diverse Semantics. IEEE Trans Cybern 44(12):2493–2508
47. Qian X, Feng H, Zhao G, Mei T (2014) Personalized Recommendation Combining User Interest and Social Circle. IEEE Trans Knowl Data Eng 26(7):1487–1502
48. Qian X, Xue Y, Tang Y, Hou X, Mei T (2015) Landmark Summarization with Diverse Viewpoints. IEEE Trans Circuits and Syst Video Technol 25(11):1857–1869
49. Qian X, Zhao Y, Han J (2015) Image Location Estimation by Salient Region Matching. IEEE Trans Image Process 24(6):4348–4358
50. Qian X, Wang H, Zhao Y, Hou X, Hong R, Wang M, Tang YY (2017) Image Location Inference by Multisaliency Enhancement. IEEE Trans Multimedia 19(4):813–821
51. Qian X, Lu D, Wang Y, Zhu L, Tang YY, Wang M (2017) Image Re-Ranking Based on Topic Diversity. IEEE Trans Image Process 26(8):3734–3747

52. Shen J, Meng W, Yan S, Pang H, Hua X (2010) Effective music tagging through advanced statistical modeling. SIGIR
53. Wang C, Jing F, Zhang L, Zhang H (2007) Content-based image annotation refinement. In: Proc. CVPR
54. Wang X, Zhang L, Li X, Ma W (2008) Annotating images by mining image search results. IEEE Trans Pattern Anal Mach Intell 30(11):1919–1932
55. Wang X-J, Yu M, Zhang L, Cai R, Ma W-Y (2009) Argo: Intelligent Advertising by Mining a User's Interest from His Photo Collections. ACM Data Mining and Audience Intelligence for Advertising, p 18–26
56. Wang X, Zhang L, Liu M, Li Y, Ma W (2010) ARISTA - image search to annotation on billions of web photos. In: Proc. CVPR, p 2987–2994
57. Wang M, Yang K, Hua X, Zhang H (2010) Towards a Relevant and Diverse Search of Social Images. IEEE Trans Multimedia 12(8):829–842
58. Wang M, Ni B, Hua X, Chua T (2012) Assistive Tagging: A Survey of Multimedia Tagging with Human-Computer Joint Exploration. ACM Comput Surv 44(4)
59. Wang M, Hong R, Li G, Zha Z, Yan S, Chua T (2012) Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification. IEEE Trans Multimedia 14(4):975–985
60. Wang M, Li H, Tao D, Lu K, Wu X (2012) Multimodal Graph-Based Reranking for Web Image Search. IEEE Trans Image Process 21(11):4649–4661
61. Weinberger K, Slaney M, van Zwol R (2008) Resolving tag ambiguity. In: Proc. ACM Multimedia, p 111–119
62. Wu L, Hua X-S, Yu N, Ma W-Y, Li S (2008) Flickr distance. In: Proc. ACM Multimedia, p 31–40
63. Wu L, Yang LJ, Yu NH, Hua XS (2009) Learning to Tag. In: Proc. of ACM WWW
64. Xu H, Wang J, Hua X, Li S (2009) Tag refinement by regularized lda. In: Proc. ACM Multimedia
65. Yan Y, Nie F, Li W, Gao C, Yang Y, Xu D Image classification by cross-media active learning with privileged information. IEEE Trans Multimedia 18(12):2494–2502
66. Yang K, Hua X, Wang M, Zhang H (2011) Tag Tagging: Towards More Descriptive Keywords of Image Content. IEEE Trans Multimedia 13(4):662–673
67. Yang Y, Wu F, Nie F, Shen H, Zhuang Y, Hauptmann A (2012) Web and Personal Image Annotation by Mining Label Correlation with Relaxed Visual Graph Embedding. IEEE Trans Image Process 21(3):1339–1351
68. Yang Y, Nie F, Xu D, Luo J, Zhuang Y, Pan Y (2012) A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. IEEE Trans Pattern Anal Mach Intell :723–742
69. Yang X, Qian X, Xue Y (2015) Scalable Mobile Image Retrieval by Exploring Contextual Saliency. IEEE Trans Image Process 24(6):1709–1721
70. Yang Y, Ma Z, Nie F, Chang X, Hauptmann AG Multi-class active learning by uncertainty sampling with diversity maximization. Int J Comput Vis 113(2):113–127
71. Yang Y, Ma Z, Hauptmann AG, Sebe N Feature selection for multimedia analysis by sharing information among multiple tasks. IEEE Trans Multimedia 15(3):661–669
72. Yuan Y, Zheng X, Lu X (2017) Hyperspectral Band Selection by Discovering Diverse Subset in Multiple Graphs. IEEE Trans Image Process 26(1)
73. Zha Z, Hua X, Mei T, Wang J, Qi G, Wang Z (2008) Joint multi-label multi-instance learning for image classification. In: Proc. CVPR
74. Zha Z, Wang M, Zheng Y, Yang Y, Hong R, Chua T (2012) Interactiv e Video Indexing With Statistical Active Learning. IEEE Trans Multimedia 14(1):17–27
75. Zhang S, Huang J, Li H, Metaxas D (2012) Automatic Image Annotation and Retrieval Using Group Sparsity. IEEE Trans Syst Man Cybern Part B Cybern 42(3):838–849
76. Zhang D, Han J, Jiang L, Ye S, Chang X (2017) Revealing Event Saliency in Unconstrained Video Collection. IEEE Trans Image Process 26(4):1746–1758
77. Zhao G, Qian X, Lei X (2016) Objective Evaluation for Service by Deep Exploring Social Users' Contextual Information. IEEE Trans Knowl Data Eng 28(12):3382–3394
78. Zhao G, Qian X, Xie X (2016) User-Service Rating Prediction by Exploring Social Users' Rating Behaviors. IEEE Trans Multimedia 18(3):496–506
79. Zhao G, Qian X, Kang C (2017) Service Rating Prediction by Exploring Social Mobile Users' Geographical Locations. IEEE Trans Big Data 3(1):67–78
80. Zhou N, Cheung WK, Qiu G, Xue X (2011) A Hybrid Probabilistic Model for Unified Collaborative and Content-Based Image Tagging. IEEE Trans Pattern Anal Mach Intell 33(7):1281–1294

Miao Shen received the B.S. degrees in Xi'an University of Technology, Xi'an, China, in 1999. He was an engineer at Xi'an Jiaotong University City College from Nov. 2008