reprinted from: IPSJ Transactions on Databases 6 (3) : 177–186 (2013)
IPSJ Online Transactions Vol.6 75–84 (July 2013)
[DOI: 10.2197/ipsjtrans.6.75]

Regular Paper

Tag Quality Improvement for Social Image Hosting Website

Jiyi Li[1,a)] Qiang Ma[1,b)] Yasuhito Asano[1,c)] Masatoshi Yoshikawa[1,d)]

Received: December 20, 2012, Accepted: April 1, 2013, Released: July 4, 2013

Abstract: Social image hosting websites such as Flickr provide services to users for sharing their images. Users can upload and tag their images or search for images by using keywords which describe image semantics. However various low quality tags in the user generated folksonomy tags have negative influences on image search results and user experience. To improve tag quality, we propose four approaches with one framework to automatically generate new tags, and rank the new tags as well as the existing raw tags, for both untagged and tagged images. The approaches utilize and integrate both textual and visual information, and analyze intra- and inter- probabilistic relationships among images and tags based on a graph model. The experiments based on the dataset constructed from Flickr illustrate the effectiveness and efficiency of our approaches.

Keywords: tag quality, social image hosting websites, graph model

1. Introduction

On social image hosting websites, e.g., Flickr [1], users can upload and tag their images, to share them with other users. As which has been investigated in Ref. [2], in all social tags generated by users, tags which are used to describe image content and semantics occupy the largest proportion; in all queries that users use for searching images, queries which are related to image content have the largest proportion. It is to say that many social tags can be used to index image content on semantics; and users can search images by using this kind of social tags as keywords. Although user generated tags are useful for social image management and sharing, there is a problem that they are folksonomy tags [2]. In contrast to taxonomy tags, they have an open vocabulary and very free on type, form and content.

We use **Fig. 1** as an example to illustrate various cases of low quality tags.

- *Missing Tag:* Because choosing proper tags manually for large amount of images is so time consuming, many users may miss some important tags which describe image content when they assign tags to images, e.g., the missing tag "grass" in Fig. 1. The special case of missing tag is no tag which means that users do not assign any tag to images. Missing tag causes that this image cannot appear in the corresponding search results.
- *Imprecise Tag:* For tagged images, in many cases, the assigned tags are neither precise nor meaningful enough for reflecting image semantics, e.g., the imprecise tag "white-horse" in Fig. 1. It results in that this image will appear in the incorrect keyword search results or it will not appear in the correct keyword search results.
- *Meaningless Tag:* Some tags, e.g., "D200" in Fig. 1, are not used to describe the content of an image. They are irrelevant to image semantic and meaningless for searching image content. It influences the search results accuracy and time cost.
- *Unranked Tag:* Figure 1 also shows that there is no information to describe the importance of tags and the raw tags are unranked. It causes that this image may have an inappropriate rank value in the corresponding search results.

As a result, low quality tags will decrease the search results quality; and users who want to share or search images will fail to reach their purposes.

To improve search results quality as well as user experience, one of the solutions is to improve tag quality automatically. We generate precise and meaningful tags which can reflect the objective content of images or how most of users understand the image content from a statistical viewpoint. We propose a solution with various functions which include generating new tags, and ranking these new tags as well as the existing raw tags automatically, for both untagged images and tagged images. Some work can be applied for some of these functions and solve the problems, such as tag recommendation, image annotation, tag ranking and so on.

The tag recommendation approaches using textual information only, e.g., Ref. [3], cannot work automatically and depend on the initial tag set assigned by users too much.

The image annotation approaches using visual information only [4] have been originally proposed for classifying images into a small number of concepts. Reference [8] concentrates on the underlying among different tags which are labeled to an image while straightforward methods consider these tags independently. They are proper for taxonomy keywords, but not proper enough

[1] Department of Social Informatics, Kyoto University, Kyoto 606–8501, Japan
[a)] jyli@db.soc.i.kyoto-u.ac.jp
[b)] qiang@i.kyoto-u.ac.jp
[c)] asano@i.kyoto-u.ac.jp
[d)] yoshikawa@i.kyoto-u.ac.jp

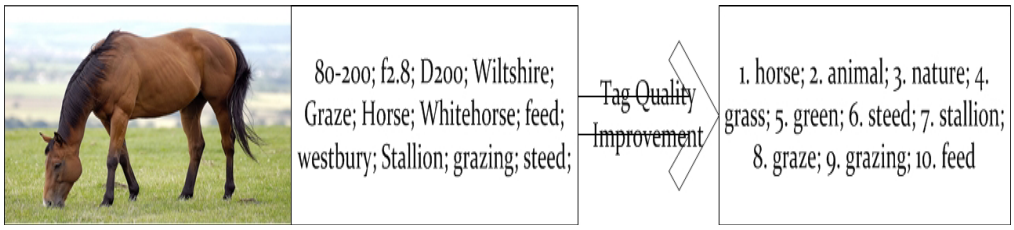

Fig. 1 Social image, raw tags, tag quality improvement.

for folksonomy tags which have a large concept space. They have a separate training stage and need to construct a training set manually which is time consuming.

Furthermore, Ref. [5] proposed by Liu et al. is related to the case of unranked tag. It ranks the raw tags of images by using a linear combination on textual and visual information with a random walk process. However, it is not related to generate new tags and ranking them with raw tags.

The contributions of our work are as follows. We improve image search results on social image hosting websites by improving tag quality. We propose a unified framework which has two stages to solve the problems. The first stage is to collect the candidate images with visual information and candidate tags with textual information. The second stage uses these information to analyze the relevance relationships between images and tags. We propose an image-tag graph model for analyzing these relationships.

We propose a series of approaches as the solutions of the second stage. Three of these approaches refer the approaches in the above mentioned existing work. In these three approaches, the first one utilizes textual co-occurrence information among all candidate tags. The second one is based on random walk method and uses visual similarity and textual co-occurrence for constructing the transition matrix. The third one aggregates the visual similarity of the related images of a tag. With the usage of the unified framework and our modification, the disadvantages of these approaches can be solved, and they can handle above-mentioned functions to meet various cases of low quality tags.

In addition, we also propose an original approach which mixes the textual and visual information. It has a mutual reinforcement process to propagate these information through the graph edges. All these approaches do not have the disadvantage of too depending on the initial tag set. We improve tag quality based on user generated folksonomy tags directly, which have a large concept space, without constructing and using manual training set.

The experimental results show that all these approaches can improve tag quality prominently, while in contrast to the approaches adapted from existing work, our original approach has better performance on the NDCG metric and time cost. On the other hand, our approaches can also be used for the task of ranking raw tags only without adding new tags. For this task, all our approaches can improve tag quality. It also shows that for improving tag quality, compared with ranking raw tags only, adding new tags has better performance on the MAP metric.

The remainder of this paper is organized as follows. In section 2, we introduce the related work. Section 3 presents our approaches for improving tag quality. In section 4, we report the experimental results. Section 5 presents the conclusion.

2. Related Work

We concentrate on the related work on automatic tag recommendation for images since they are most related to our work. The approaches proposed in this area can be divided into several categories based on the information they use, i.e., text-based approaches and visual-based approaches.

The text-only-based approaches only use textual information for tag analysis. Reference [3] is a typical text-only-based approach. It uses two tag co-occurrence measures, and aggregate three types of tags with two strategies. This approach which only uses textual information has the deficits that it cannot work automatically and depends on the initial tag set. It assumes that a user assigns a few candidate tags to the input image manually. It cannot recommend tags to an untagged image automatically. It also does not have a good characteristics of fault tolerance on the initial tag set. A fault initial tag set will be certain to generate fault result. Furthermore for different images with same initial tag sets, it will generate same results. It cannot generate diverse enough results for different images. Our approaches, because of the usage of visual information, do not so depend on the initial tag set and can handle untagged images as well as tagged images. They can generate better results even if the initial tag set is wrong, and generate different results for different images even the initial tag sets are same.

The visual-only-based approaches only use image content for tag analysis. Reference [6] uses an image annotation approach. It defines and learns 62 concepts, and annotates new images with the top-n concepts. It has a learning process and the size of concepts space is fixed and small. The approaches only using visual information have been originally proposed for classifying images into a small number of concepts which can be regarded as tags. They need to create the training set manually which costs lots of time and labor. Adding a new concept for the classification takes computational time for reconstructing the classifiers. Such approaches are not suitable for user generated folksonomy tags. Furthermore the performance of content-based approaches for image retrieval nowadays is still not better than text-based approaches. In our work we propose approaches which utilize user generated tags directly, without training set construction manually and training process. They make it possible and easier to handle large concept space. We also integrate textual information with visual information.

Several approaches using both textual and visual information have also been proposed. Reference [9] focuses on a more precise

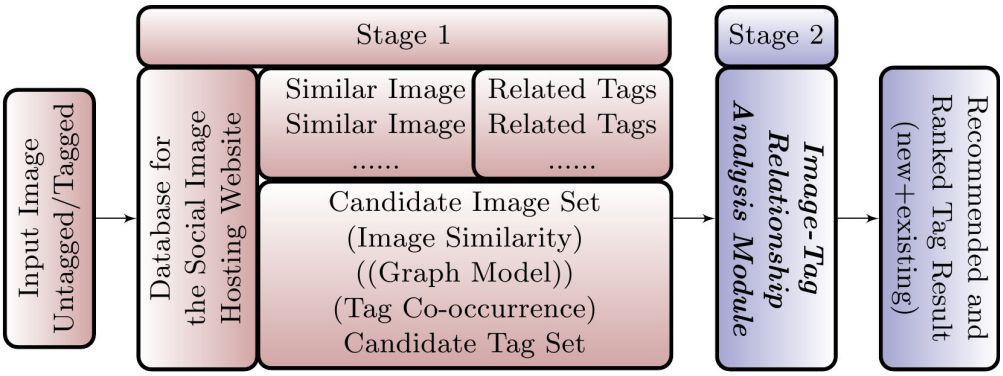

Fig. 2 Overview of framework.

description of the multimedia object by exploring the information capability of individual tag and the tag-to-set correlation. It aggregates the visual similarity of the related images to measure the relevance between a tag and an image. Reference [7] generates a ranking feature for tags with textual and visual modality respectively, and then use Rankboost algorithm to learn an optimal combination. Reference [5] provides a linear combination on textual and visual information with a random walk process. Reference [10] uses textual information to generate candidate tags and leverages visual information to filter noisy tags with clustering method. These approaches combine these two kinds of information. The original approach we propose mixes both textual and visual information with each other.

3. Tag Quality Improvement

In this section we first propose the framework of our approaches for improving the tag quality. After that we introduce the graph model constructed for relationship analysis among images and tags. At last we propose the adapted approaches and our original approach in details respectively.

3.1 Approach Framework

The relevance relationships between images and tags can be presented in probability. In our work, for an input image q of which we will improve the tag quality, a tag t has a probability of $p(t|q)$ to represent its semantics. This probability depends on the textual information $c(t)$ based on this tag and visual information $s(q)$ based on this query: $p(t|q) = f(c(t), s(q))$.

Figure 2 shows the framework of our approaches. It has two main stages.

Stage 1 (Candidate Set Construction): Given an input image q, We find the top-k (our current work, $k = 100$) similar tagged images in the database $\mathcal{D}$ and construct a candidate image set A with them. We construct a candidate tag set T with user generated tags of these images. For image a_i, its own tag set is T^{a_i}. The similarity between q and a_i is s_i. We denote t_u as a tag.

Stage 2 (Relevance Relationship Analysis): We analyze the image-tag relationship on candidate set A and T. After that, we can get the relevance probability of each tag to q, and finally tag rank list for q. We do not update the original database after this stage immediately. The processing of the next input image is not influenced by the improved tag set of the current input image. After all of the images have been processed, we get a tag quality improved database.

The advantages of this framework with two stages can be addressed into two dimensions. On one hand, although both of them will influence the final results, these two stages is relatively independent with each other. Stage 1 is to collect the candidate image and tag information, Stage 2 is to analyze the relationships of information. It is easier to optimize the task of tag quality improvement at different stages separately. In this paper, we do not concentrate on how to improve the performance of Stage 1, such as proposing optimized method of feature extraction and top similar images search. Many methods from the area of content based image retrieval can be utilized in Stage 1. The performance of Stage 1 can be improved by the community of computer vision. We concentrate on how to evaluate the relationships among images and tags based on given candidate image and tag sets.

On the other hand, this framework makes our approaches not too depending on the initial tag set of the given input image from several aspects. First, because it uses the top similar images and their user generated tags to construct the candidate image and tag set, it can solve the problem of handling untagged images. Second, we consider all of the tags in the candidate tag set T. It makes the approaches still available even if the initial tag set of the given image is wrong. Third, for different images with the same initial tag set, it can generate different results because the top similar images of these different images are different.

In this paper, we propose four different approaches with this framework. All of them are same in Stage 1, while they are different in Stage 2. These approaches are based on the following graph model in the image-tag relationship analysis.

3.2 Image-tag Relationship Analysis Model

Figure 3 shows the graph model we proposed. It includes both intra- and inter- relationships among images and tags. It is composed of several parts, image similarity graph, tag complete graph and image-tag bipartite graph. First, q denotes current input image which can be untagged or tagged images, a_i denotes one of its top similar images. The links among images denote image similarity. For q, there is also a link point to itself, the similarity it denotes is equivalent to 1. These images and links construct the image similarity graph. Second, t_u denotes a tag. The links among tags denote tag co-occurrence. These tags and links con-

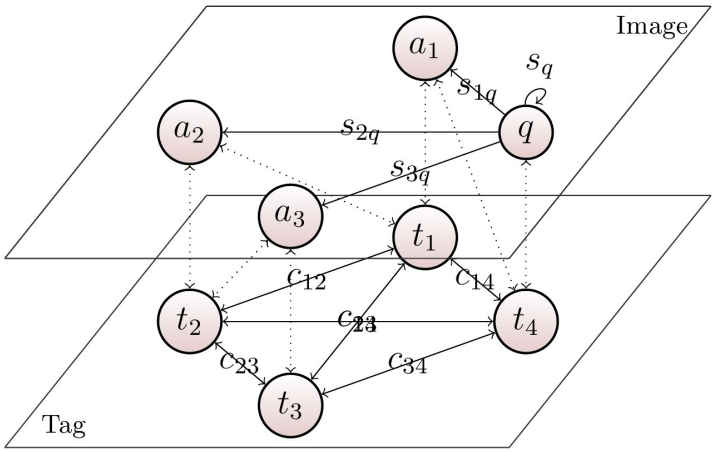

Fig. 3 Image-tag relationship analysis model.

struct the tag complete graph. Third, the links between images and tags denote image-tag annotation relationships and construct an image-tag bipartite graph.

The visual parameters s and textual parameters c in the model are computed as follows. We use visual image similarity s_i ranging from 0.0 to 1.0 as visual parameters. The textual parameters c are different for different approaches. We will introduce them in each approach. All textual parameters for all tags are computed in advance.

3.3 Existing Approach Adaption

We propose several approaches for Stage 2. Three of them are adapted from the existing work. With the framework we proposed and our modification, we can solve the disadvantages of these approaches so that they can handle all the functions we focus on and meet various kinds of low quality tags.

3.3.1 MCKA: Modified Collective Knowledge Approach

First we propose an approach that only uses textual information in Stage 2. It is only based on the tag complete graph in the model in Fig. 3. For each tag t_u in the candidate tag set T, we find the top-k most co-occurring tags t_v in T to construct the top co-occurrence tag list T_{uk}. We compute the score for each t_v and aggregate them. To promote the better tags, we also uses three *promotion* functions, *rank*, *descriptive* and *stability*, which are based on the tag frequency and tag position in the top co-occurring tag list, to generate the final results. These *promotion* functions and their value of the parameters follow the definition in Ref. [3]. The algorithm is list in Algorithm 1. The definition of the *promotion* functions are also shown in Algorithm 1. Because we consider all tags in the candidate tag set T to find their top-k most co-occurring tags, this approach is still available even if the initial tag set of the given image is wrong.

A brief description of the formulas is as follows. Here $P(t_v)$ is used to evaluate the relevance of a tag to the given image at last.

$$P(t_v) = \sum_{t_u \in T} vote(t_u, t_v) * promotion(t_u, t_v),$$

$$vote(t_u, t_v) = \begin{cases} 1 \; if t_v \in T_{uk} \\ 0 \text{ otherwise} \end{cases},$$

$$promotion(t_u, t_v) = rank(t_u, t_v) * descriptive(t_v) * stability(t_u).$$

Algorithm 1 Modified Collective Knowledge Approach

```
Input: Candidate Image Set A, Candidate Tag Set T
Output: Rank List score for T
Initial: zeros(score)
for all t_u ∈ T do
    stability(t_u) = k_s/(k_s + abs(k_s − log|t_u|))
    C(u,T) = {c(u,v)|t_v ∈ T}
    T_uk = {t_v|c(u,v) ∈ Topk(C(u,T))}
    for t_v ∈ T_uk do
        rank(t_u,t_v) = k_r/(k_r + (r − 1)), r = the position of t_v in T_uk
        P(t_v)+ = stability(t_u) * rank(t_u,t_v)
    end for
end for
for all t_v ∈ T do
    P(t_v)* = descriptive(t_v), descriptive(t_v) = k_d/(k_d + abs(k_d − log|t_v|))
end for
```

The textual parameter $c(u, v)$ of each pair of tag t_u and t_v is computed to evaluate their co-occurrence degree. In Ref. [3], Sigurbjörnsson et al. refer two measures. One is a asymmetric measure $c(u, v)_1$. The other is a symmetric measure $c(u, v)_2$.

$$c(u,v)_1 = \frac{|t_u \cap t_v|_{\mathcal{D}}}{|t_v|_{\mathcal{D}}}, \quad c(u,v)_2 = \frac{|t_u \cap t_v|_{\mathcal{D}}}{|t_u \cup t_v|_{\mathcal{D}}}.$$

$|t_u|_{\mathcal{D}}$ means the number of images that contain t_u in the database, $|t_u \cap t_v|_{\mathcal{D}}$ means the number of the images that contain both of t_u and t_v. $|t_u \cup t_v|_{\mathcal{D}}$ means the number of the images that contain t_u or t_v. We propose a symmetric-asymmetric pairwise co-occurrence parameter for this approach:

$$c(u,v) = \frac{|t_u \cap t_v|_{\mathcal{D}}}{|t_u|_{\mathcal{D}}} + \frac{|t_u \cap t_v|_{\mathcal{D}}}{|t_v|_{\mathcal{D}}}.$$

3.3.2 RWBA: Random Walk Based Approach

MCKA only uses textual information in Stage 2. We therefore propose an approach that uses both visual information with textual information in the image-tag relationship analysis. The iteration process of this approach is mainly based on the tag complete graph in the model in Fig. 3. It also uses the information of image similarity relationship and image-tag annotation relationship. It utilizes the random walk method in the iteration process. We use image similarity information as well as image-tag annotation relationship to construct the transition matrix, and integrate tag co-occurrence information in the iteration process.

This approach is based on an assumption. For an input image, a high quality tag has a high co-occurrence with other high quality tags. Suppose that when we want to generate tags for an image, if we have confirmed a tag t and then we want to generate more tags, the tags which is most associated with t will be the good candidates and we will assign them with high relevance probability.

We initial a tag score with a tag single co-occurrence parameters:

$$c_u = \sum_v c(u, v).$$

$c(u, v)$ is same with the one in MCKA. The iterative formulas for computing the score of a tag are as follows.

$$P_k(t_u) = \gamma P'(t_u) + (1 - \gamma) \sum_v (P_{k-1}(t_v) P(t_v \to t_u)),$$

$$P'(t_u) = \frac{c_u}{\sum_v c_v},$$

$$P(t_v \to t_u) = \frac{r(v, u)}{\sum_w r(v, w)},\ r(v, u) = e^{-\frac{\sum_{t_v \in T^{a_i}} \sum_{t_u \in T^{a_j}} \|s_i - s_j\|}{|t_v|_T * |t_u|_T}}.$$

$P'(t_u)$ denotes the probability that we may not turn to recommend other tags and keep the state at current tag. The state transition matrix $R = \{P(t_v \to t_u)\}_{m*m}$ is decided by visual similarity information and image-tag annotation relationship. a_i is an image in the candidate image set A, and T^{a_i} is the corresponding tag set of image a_i. $|t_v|_T$ is the number of images in A that are pointed to by tag t_v.

This approach is similar but different from tag ranking approach [5]. Both RWBA and tag ranking use the random walk method. But they have the following difference. First, the tag ranking approach is originally for ranking the existing raw tags of the images. After the database has been modified into a database with the tags ranked, it applies this updated database for tag recommendation as an application of their work. The tag recommendation of their work no longer uses random walk method and just uses the tag co-occurrence. Second, the detailed computation of these two approaches such as initialization and transition matrix are different.

3.3.3 ISAA: Image Similarity Aggregation Approach

We propose this approach by referring the idea from a state-of-art approach in the area of tag recommendation, Ref. [9]. In this work, it aggregates the visual similarity of the related images to measure the relationship between a tag and a given image. For each image a_i in the top-k similar tagged images, we first evaluate their semantic consistency $Cs(a_i)$ by using

$$Cs(a_i) = \frac{1}{k} \sum_{\vee a_j : a_j \in A} d(T^{a_i}, T^{a_j}) * s_j,$$

where $d(T^{a_i}, T^{a_j})$ is the cosine distance of two tag set T^{a_i} and T^{a_j}. After that, we select top-r semantic consistent images $\mathcal{A}_r$ from these images for further computation. The rules of our selection is

$$Cs(a_i) > min_j\{Cs(a_j)\} + (max_j\{Cs(a_j)\} - min_j\{Cs(a_j)\}) * \eta.$$

At last, we evaluate the relevance score of the candidate tags by the following formula.

$$P(t_u) = \frac{\sum_{\vee t_u : t_u \in T^{a_i}, a_i \in \mathcal{A}_r} s_i}{r}.$$

3.4 Our Original Approach

In addition to the approaches adapted from existing work, we also propose an original approach of image-tag relationship analysis. In contrast to the above approaches, our original approach more effectively mixes the textual and visual information with each other by propagating the information through the graph edges with a mutual reinforcement process.

3.4.1 MMRA: Mixed Mutual Reinforcement Approach

The iteration process of this approach is mainly based on the image-tag bipartite graph, which denotes the inter- relationships of the graph model in Fig. 3. The candidate image set and tag set construct two disjoint sets of the bipartite graph. This approach also considers the information on tag complete graph and image similarity graph which denote the intra- relationships. MMRA is based on a basic assumption: a high quality tag for q is a tag that point to many high quality images; a high quality image is an image that is pointed to by many high quality tags. A high quality tag means this tag has high relevance probability to be a tag of the input image. A high quality image can be regarded as an image that has high semantic similarity with the input image. We initiate the score $Q(t)$ of the tags with the textual parameters and the score $Q(a)$ of the images with the image parameters. We will discuss this textual parameter later.

Initialization: $Q'_0(a_i) = \Phi(s_i),\ Q'_0(t_u) = \Phi(c_u)$;

After the initialization, we have an iterative computation to compute $Q(t)$ and $Q(a)$. When we design the iteration formulas, we need to make it following the requirements of the above basic assumption. A normalization is also necessary to solve the convergence problem of this iterative computation. Based on these conditions of designing the approach, we come to the following iterative formulas.

Iteration:

$$\begin{cases} Q_{k+1}(t_u) = \alpha\Phi(c_u) + (1 - \alpha) \sum_{\vee a_i : t_u \in T^{a_i}} \Phi(s_i) Q'_k(a_i) \\ Q_{k+1}(a_i) = \beta\Phi(s_i) + (1 - \beta) \sum_{\vee t_u : t_u \in T^{a_i}} \Phi(c_u) Q'_k(t_u) \\ Q'_{k+1}(a_i) = \Phi(Q_{k+1}(a_i)),\ Q'_{k+1}(t_u) = \Phi(Q_{k+1}(t_u)) \end{cases}$$

$$0 \le \alpha, \beta \le 1$$

$$\Phi(Q_k(t_x)) = \frac{Q_k(t_x) - min_y\{Q_k(t_y)\}}{max_y\{Q_k(t_y)\} - min_y\{Q_k(t_y)\}}.$$

Here "$\vee a_i : t_u \in T^{a_i}$" means for all of the image that are pointed by tag t_u, "$\vee t_u : t_u \in T^{a_i}$" means for all of the tags that pointed to image a_i. The iteration parameters α and β are damping factors. k is the number of iteration steps. $\Phi(\cdot)$ a Max-Min normalization function, we use $Q_k(t_x)$ as an example to illustrate its definition. With the observation of the intermediate results, we find that the iteration can always converge in several iterations. We therefore set a fixed maximum of iterations as 10.

Visual image similarity to the given image is an inherent property of a candidate image. The images which have high similarity

can be regarded as more important on the graph. A similar property is also observed for a candidate tag. We therefore use visual parameters and textual parameters as the weights of images and tags in the iterations. These weights represent the importance of these images and tags on the graph.

3.4.2 Textual Parameter

The textual parameter needs to reflect both the local information in the candidate tag set and the global information in the whole dataset. We therefore define two frequency metrics: local frequency $|t_u|_{\mathcal{T}}$, which is defined as the number of images with tag t_u in the candidate image set, and global frequency $|t_u|_{\mathcal{D}}$, which is defined as the number of images with tag t_u in the database. A naive definition of the textual parameter is $c_u = |t_u|_{\mathcal{T}}/|t_u|_{\mathcal{D}}$. However, this parameter is not strong enough to handle some cases. For example:

Case 1: If there is a good tag t_u with $|t_u|_{\mathcal{T}} = 10$ and $|t_u|_{\mathcal{D}} = 1500$; another good tag t_v with $|t_v|_{\mathcal{T}} = 5$ and $|t_v|_{\mathcal{D}} = 750$. In such case, c_u is equal to c_v. These two tags have same value on the parameter. However, t_u should be more important than t_v in the candidate set.

Case 2: If there is a good tag t_u with $|t_u|_{\mathcal{T}} = 10$ and $|t_u|_{\mathcal{D}} = 1500$; another noisy tag t_v with $|t_v|_{\mathcal{T}} = 1$ and $|t_v|_{\mathcal{D}} = 100$. In such case, $c_u < c_v$. The noisy tag has higher value on the descriptor.

To solve the problem of case 1, one solution is to propose a more reasonable descriptor, another solution is to leverage the iteration process. In the iteration process, for case 1, even if c_u is equal to c_v, t_u still has larger influence than t_v on the graph and in the iterations. It is because in an iteration, for all images, $\Phi(c_u)Q'_k(t_u)$ will be computed 10 times, $\Phi(c_v)Q'_k(t_v)$ will be computed only 5 times. More scores from t_u will be propagated through the graph edges.

To solve the problem of case 2. On one hand, which is similar to the solution in case 1, our iteration process can somewhat handle the problem. On the other hand, we regard the tags with too low local frequency as noisy tags and filter them.

At last, the textual parameter for our original approach is defined as follows:

$$c_u = \begin{cases} \frac{|t_u|_{\mathcal{T}}}{|t_u|_{\mathcal{D}}}, & if\ |t_u|_{\mathcal{T}} > \delta, \\ 0, & if\ |t_u|_{\mathcal{T}} \le \delta. \end{cases}$$

4. Experiments

4.1 Experimental Settings

We set several rules to filter the raw tag set before computing textual parameters. We delete the space in tags and convert them into lower case format so that these tags can be regards as a single tag. We eliminate the tags that are not existing in WordNet, because we consider that most of them are misspelled or irregular tags and have little contribution to the keyword search. We also eliminate the numeric tags. Furthermore, we eliminate the tags with too low frequency, which are misspelled tags, too special tags and so on. This processing can reduce the noises in the tag set and the time cost in the computation of image-tag relationship analysis. We do not eliminate the tags with too high frequency. Although lots of previous work think that this kind of tags is too common to represent the semantic of content. We think that this kind of tags can reflect important semantic and user tagging behavior with image categories.

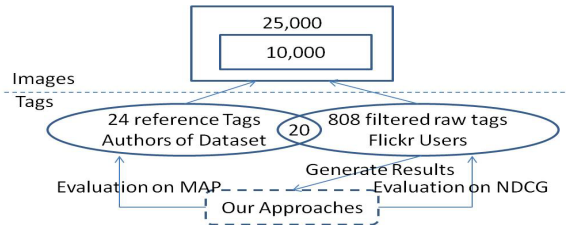

Fig. 4 Data and evaluation.

The dataset we use for experiment is MIR Flickr 25000 [12] which is constructed by images and tags downloaded from Flickr. It has 25,000 images and provides the raw tags of these images on Flickr as well as a manual reference tag set. In this paper, we use the first 10,000 images in the dataset for evaluation. Based on our tag filter rules, we set the low frequency threshold as 20 and the setting of WordNet search as NOUN. There are 808 unique tags left after filtering.

This dataset provides image annotation ground truth on 24 concepts. We regard them as reference tag set. In our work, the version of the reference tag set we use is v080[*1]. There are 24 unique reference tags in total. Because the original motivation of this ground truth on 24 concepts is to learn with an image subset in the dataset with the labels of this ground truth and test on the other images with this ground truth. The ground truth is somewhat inconsistency with the raw tag set. 4 of 24 concepts never appear in the raw tag set, which means that it will also never appear in our results because our approaches improve tag quality utilizing the filtered user generated raw tags directly, not the reference tags. We therefore use the left 20 unique reference tags for evaluation. We also do not need to divide this dataset into training set and testing set, and do not need to train on the training set. It is different from the image annotation task on this dataset. **Figure 4** provides an intuitive presentation on these data and evaluation on them.

The visual feature we use for searching the top similar images is a 1024-Dimension color histogram feature on the HSV color space. The distance between image a_i and a_j is computed using a Pearson correlation distance $s(a_i, a_j)$ defined as

$$\mathcal{H}'_i(x) = \mathcal{H}_i(x) - \frac{\sum_y \mathcal{H}_i(y)}{\mathcal{N}},$$

$$s(a_i, a_j) = \frac{\sum_x (\mathcal{H}'_i(x) * \mathcal{H}'_j(x))}{\sqrt{(\sum_y \mathcal{H}'_i(y)^2) * (\sum_y \mathcal{H}'_j(y)^2)}},$$

where $\mathcal{H}_i$ and $\mathcal{H}_j$ are feature vectors. $\mathcal{N}$ is the size of the feature vector.

4.2 Metrics

We set several metrics to evaluate the statistical performance of our approaches. Each metric can be regarded as one aspect of the tag quality. If a metric is improved, the tag quality on the aspect of this metric is improved. The metrics are as follows:

MAP: The Mean Average Precision is a good metric for the

[*1] There are two sub versions for some tags in this version, e.g., flower and flower_r1. We use the "r1" version for them.

Table 1 Example of NDCG relevance degree.

	Degree	4	3	2	1	0
	Description	very relevant	relevant	partially relevant	weakly relevant	irrelevant
	Tag	flower, purple, yellow	plant	garden, nature	macro	sky

performance evaluation of the rank result in the information retrieval area. In this experiment, specially, for the raw tag set, because it originally doesn't have the rank information. Without loss of generality, we simply use the position of these tags in the raw tag list as their rank value when we compute MAP. We evaluate MAP on the whole dataset.

$$MAP = \frac{\sum_{r=1}^{N} P(r) * r}{NumberofReferenceTags}.$$

$$P(r) = \frac{|t : R(t) \leq r, t \in T_{Ref} \cap T_{Res}|}{r}$$

$rel(r) = 0$, when the tag on rank position r is not in the reference tag set; $rel(r) = 1$, when the tag on rank position r is in the reference tag set. T_{Ref} is the 20 unique reference tag set and T_{Res} is the ranked tag result. $N = max\{r(t_u)|t_u \in T_{Ref} \cap T_{Res}\}$ is the mininum of the tags in the rank results that contain all of the reference tags to this image. $r(t_u)$ is the rank value of t_u in the ranked tag result.

MPFRR and MPLRR: Mean Precision at the First Relevant Rank (PFRR) is the mean of precision at the position of the first relevant tag in the rank list on the evaluation dataset. Mean Precision at the Last Relevant Rank (PLRR) is the mean of precision at the position of the last relevant tag in the rank list on the evaluation dataset.

$$PFRR(q) = \frac{1}{min\{r(t_u)|t_u \in T_{Ref} \cap T_{Res}\}}$$

$$PLRR(q) = \frac{1}{max\{r(t_u)|t_u \in T_{Ref} \cap T_{Res}\}}$$

NDCG: Normalized Discounted Cumulative Gain (NDCG) [13]. We evaluate the approaches on the metrics of MAP, MPFRR and MPLRR based on the ground truth on 20 unique reference tags. However, in this ground truth, most of images can only be labeled with quite few tags, such as zero or one tag. It is limited to reflect the performance of tag quality improvement on the large amount of images and their tags in the dataset. We therefore construct another ground truth for evaluation on NDCG, which evaluates the rank results of all tags in the candidate tag set. NDCG is an effective metric often used in information retrieval for evaluating the rank results with relevance levels. For a given result, when more tags with higher relevance scores are ranked higher, the NDCG score of this result is higher. It is defined as follows,

$$NDCG@k = Z_k \sum_{j=1}^{k} \frac{2^{rel(j)} - 1}{log(1 + j)}.$$

$rel(j)$ is the relevance level of the tag at rank j. Z_k is a normalization constant and equal to the maximum DCG value that the top-k ranked tags in the whole tag list can reach, so that NDCG score is equal to 1 for the optimal results of which the relevance scores have a descending order. We randomly select 100 images, a subset of the dataset, and label the relevance degrees of all tags in the candidate tag sets of these images on the HSV feature by human beings. In total, three people take participate in this labeling task. The range of relevance degree is from 0 (irrelevant) to 4 (relevant). We evaluate average NDCG value on this subset of the dataset. **Table 1** shows an intuitive example of different relevance levels.

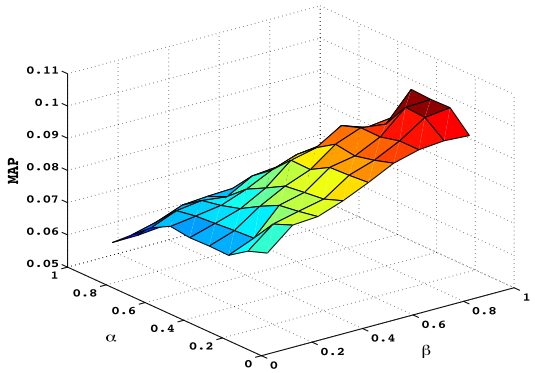

Fig. 5 Parameters for MMRA.

Table 2 δ of textual parameter.

δ	0	1	2	3	4
MAP	0.0832	0.0884	0.1016	0.1051	0.1075

4.3 Approach Parameters

We set the parameters of the approaches as follows. For MCKA, $k = 25$, and the parameters in the promotion functions follow the value in Ref. [3], $ks = 9$, $kd = 11$, $kr = 4$. For RWBA, the value of the damping factor parameter γ follows the general choice of the damping factor in PageRank [11], which can also be regarded as a random walk based method, $\gamma = 0.15$. For ISAA, the parameter η for selecting semantic consistent images is set to $\eta = 0.15$.

For MMRA, to decide the iteration parameters of α and β, we choose their candidate values by an interval of 0.1 in the range of (0.1, 0.9) and get 81 groups of candidate values. We run the MMRA with these candidate parameter groups on first 100 images in the dataset, based on the HSV feature, and observe the performance on the metric of MAP. According to **Fig. 5**, we choose (α, β) as (0.2, 0.8) in our following experiment to evaluate the performance of this approach on the whole dataset and all metrics. For δ in textual parameter, we set different value to it and observe how the performance on MAP changes. **Table 2** shows the results. When the value of δ increases to 2, the performance has prominent improvement. After that, when δ continues increasing, the performance has minor improvement. We set $\delta = 2$ mainly based on the following two reasons. We don't set δ with too high value because it will increase the probability of deleting relevant tags. The best δ may also be different for different image samples or data sets.

Table 3 Statistical results on MAP, MPFRR, MPLRR.

Metric	Raw	MCKA:All	RWBA:All	ISAA:All	MMRA:All
MAP	0.029	0.147	0.155	0.112	0.100
MPFRR	0.212	0.328	0.368	0.229	0.216
MPLRR	0.028	0.039	0.032	0.035	0.032
Metric	Raw	MCKA:Raw	RWBA:Raw	ISAA:Raw	MMRA:Raw
MAP	0.029	0.064	0.065	0.061	0.059
MPFRR	0.212	0.450	0.452	0.441	0.437
MPLRR	0.028	0.070	0.069	0.069	0.069

Table 4 Statistical results on NDCG.

NDCG@	MCKA	RWBA	ISAA	**MMRA**
5	0.240	0.189	0.452	0.534
10	0.234	0.189	0.437	0.542
20	0.282	0.221	0.482	0.588
50	0.357	0.318	0.554	0.686

4.4 Experimental Results

Table 3 shows the experimental results on the metrics of MAP, MPFRR and MPLRR. Considering the columns of "Raw" and "*:All" first, all of the approaches have improved the tag quality on MAP prominently. All of the approaches can also somewhat improve tag quality on MPFRR and MPLRR. Furthermore, **Table 4** shows that MMRA performs better in our experiment on NDCG for the subset of dataset among these approaches, specially on the top-ranked tags.

Although our approaches are originally designed to recommend new tags, and rank the new tags as well as the existing raw tags. They are very easy to be applied to the task of ranking raw tags only, by extracting and sorting the raw tags according to their rank sequence in the whole candidate tag set directly. Note that it's not a pure "ranking raw tag only" results. When we rank these raw tags, we also utilize the information of other new tags. Table 3 also provides the result on this issue. Comparing the columns of "Raw" and "*:Raw", we can find that only ranking the raw tags can also improve tag quality from on MAP, MPFRR, MPLRR. The "*:Raw" results for all four approaches are similar.

We also investigate another issue that if we only rank the existing raw tags without including new tags, how the tag quality can be improved and what's the difference between including new tags and not. Comparing the columns of "*:Raw" and "*:All", we can find that recommending new tags and ranking them as well as the raw tags always has better tag quality on MAP metric, but worse on MPFRR and MPLRR metrics because more irrelevant tags are included.

Table 5 illustrates some sample results for each approach. From the examples, we can find that MCKA and RWBA prefer to tags with high frequency, while the 20 unique reference tags are high frequency; if MCKA and RWBA match a reference tag in the 20 reference tag, e.g., sky, the average performance on MAP can be very high. This is one reason of the performance difference on the evaluation on MAP and NDCG. MMRA and ISAA prefer to generate diverse and rational tag list, while ISAA can somewhat be regarded as a variation of the first iteration of MMRA, but without additional iterations.

4.5 Time Complexity

Table 6 provides an overview of the time complexity comparison. Here we just consider the time complexity of image-tag relationship analysis of Stage 2 in Fig. 2, because for these approaches, the time cost of other stages are the same. In this table, n is the size of the candidate tag set, and m is the size of the candidate image set. MMRA and RWBA need an iteration process. For MMRA, which is based on the bipartite graph, for each tag, the iteration computation uses the information of the annotation relationships among images and tags. For RWBA, which is based on the tag complete graph, for each tag, the computation uses the information that refers to all of the other tags. Actually the time complexity for the iteration process of MMRA and RWBA are both $O(eI)$, where e is the number of links need to be analyzed in the graphs, and I is the iteration times. In MMRA the e_{MMRA} is equal to μmn, μ is a positive number but much smaller than 1. In RWBA, e_{RWBA} is equal to $(n-1)^2$ (for each pair of tags, two edges on two directions) because of the tag complete graph. For example, for the sample image which is used for computing and illustrating the running time in Table 6 (Our experiment environment is on Ubuntu with CPU Intel Core i7 920.). It has $m = 100$ and $n = 599$, and its $e_{MMRA} = 963$ and $e_{RWBA} = 357604$. Both of their iteration times are less than ten times. We can get that $O(\mu mnI) \ll O(n^2 I)$ and the MMRA is much faster than the RWBA.

MMRA doesn't need additional computation in the initialization step. RWBA needs to compute the transition matrix which depends on the candidate tag and image set. Even RWBA uses a fixed transition matrix so that there is no additional computation in the initialization, it still costs much more time than MMRA. MCKA does not have an iteration process. It needs to select top k most co-occurring tags for each tag in the candidate set. Furthermore, for ISAA, although it has $\epsilon mn < \mu mnI$, the computation on selecting semantic consistent images π costs some running time. In total, the MMRA has better time complexity than other three approaches. The sequence on time cost of these approaches is RWBA > MCKA > ISAA > MMRA.

In conclusion, for our experiments, all of these approaches we proposed can provide a prominent improvement on tag quality, for ranking raw tags only or for adding new tags and ranking new tags as well as raw tags. In contrast to the three approaches which are adapted from existing work, our original approach MMRA has better performance on the NDCG metric and time cost. It shows that MMRA can mix and leverage the textual and visual information more effectively. Furthermore, for the task of ranking raw tags only, these four approaches have similar performance. Adding new tags has a better performance than ranking raw tags only, on the MAP metrics, but worse on the metrics of MPFRR and MPLRR.

Table 5 Sample results.

	Raw Tag	(no tags)
	MCKA	1.blue;2.sky;3.red;4.nature;5.light;6.water;7.green;8.canon;9.macro;10.white;
	RWBA	1.sky;2.blue;3.canon;4.water;5.red;6.light;7.nature;8.night;9.green;10.white;
	ISAA	1.red;2.orange;3.macro;4.yellow;5.flower;6.sunset;7.sky;8.nature;9.abstract;10.light;
	MMRA	1.purple;2.flower;3.garden;4.abstract;5.plant;6.bee;7.yellow;8.orange;9.macro;10.red;
	Raw Tag	fishermenswharf; hdr; hdri; heiwa4126; japan; photomatix; tokyo; toyosu; geotagged ;geo:lat=356522108; geo:lon=1397675686;
	MCKA	1.sky;2.water;3.blue;4.canon;5.nature;6.city;7.sunset;8.night;9.architecture;10.landscape;
	RWBA	1.sky;2.blue;3.canon;4.water;5.nature;6.night;7.white;8.interestingness;9.city;10.sunset;
	ISAA	1.sunset;2.sky;3.water;4.sunrise;5.uk;6.reflection;7.nature;8.sea;9.beach;10.sun;
	MMRA	1.sunset;2.sunrise;3.sea;4.tokyo;5.japan;6.uk;7.sky;8.reflection;9.nature;10.beach;

Table 6 Time complexity comparison.

Approach	Time Complexity		Time Cost(s)
	Initialization	Iteration	
MCKA	$O(kn^2)$		0.62
RWBA	$O(n^2)$	$O(n^2 I)$	7.22
ISAA	$O(\pi + \epsilon mn)$		0.15
MMRA	0	$O(\mu mnI)$	0.09

5. Concluding Remarks

In this paper, we improve user generated folksonomy tag quality on social image hosting websites, to improve social image search results as well as user experience. For this topic, we propose an original approach as well as three different approaches adapted from existing work, within a unified framework and image-tag relationship graph model. The computational experiments reveal that our approaches are effective and efficient.

For future work, we would like to make further research on which kind of elements have important influences on the results, and how they influence. For example, how Stage 1 influences the final results; when using other image feature or content-based image search methods for collecting top-k similar images, how the results change; what is a proper value of k for the top-k visual similar images; how many top ranked tags we select as the final tag list to a given image is proper.

Acknowledgments This work was partly supported by JSPS KAKENHI Grant Number 25700033 and Grant-in-Aid for Young Scientists (B) Grant Number 23700116.

References

[1] Flickr: available from ⟨http://www.flickr.com⟩.
[2] Bischoff, K., Firan, C.S., Nejdl, W. and Paiu, R.: Can all tags be used for search?, *Proc. 17th ACM Conference on Information and Knowledge Management* (*CIKM'08*), pp.193–202 (2008).
[3] Sigurbjörnsson, B. and van Zwol, R.: Flickr Tag Recommendation based on Collective Knowledge, *Proc. 17th International Conference on World Wide Web* (*WWW'08*), pp.327–336 (2008).
[4] Datta, R., Joshi, D., Li, J. and Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age, *ACM Computing Surveys* (*CSUR*), Vol.40, No.2, pp.1–60 (2008).
[5] Liu, D., Hua, X.S., Yang, L.J., Wang, M. and Zhang, H.-J.: Tag Ranking, *Proc. 18th International Conference on World Wide Web* (*WWW'09*), pp.351–360 (2009).
[6] Chen, H.M., Chang, M.H., Chang, P.C., Tien, M.C., Hsu, W.H. and Wu, J.L.: SheepDog: Group and tag recommendation for flickr photos by automatic search-based learning, *Proc. 16th ACM International Conference on Multimedia* (*MM'08*), pp.737–740 (2008).
[7] Wu, L., Yang, L.J., Yu, N.H. and Hua, X.S.: Learning to Tag, *Proc. 18th International Conference on World Wide Web* (*WWW'09*), pp.361–370 (2009).
[8] Yang, Y., Wu, F., Nie, F.P., Shen, H.T., Zhuang, Y.T. and Hauptmann, A.G.: Web & Personal Image Annotation by Mining Label Correlation with Relaxed Visual Graph Embedding, *IEEE Trans. Image Processing* (*TIP*), Vol.21, No.3, pp.1339–1351 (2012).
[9] Zhang, X.M., Huang, Z., Shen, H.T., Yang, Y., and Li, Z.J.: Automatic Tagging by Exploring Tag Information Capability and Correlation, *World Wide Web*, Vol.15, Iss.3, pp.233–256 (2012).
[10] Yang, Y., Huang, Z., Shen, H.T. and Zhou, X.F.: Mining Multi-tag Association for Image Tagging, *World Wide Web*, Vol.14, Iss.2, pp.133–156 (2011).
[11] Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks and ISDN Systems*, Vol.30, Iss.1-7, pp.107–117 (1998).
[12] Huiskes, M.J. and Lew, M.S.: The MIR Flickr Retrieval Evaluation, *Proc. 1st ACM International Conference on Multimedia Information Retrieval* (*MIR'08*), pp.39–43 (2008).
[13] Jarvelin, K. and Kekalainen, J.: Cumulated gain-based evaluation of IR techniques, *ACM Trans. Information Systems* (*TOIS*), Vol.20, Iss.4 (2001).

Jiyi Li received his B.S. and M.S. in Computer Science from Nankai University, China, in 2005 and 2008, respectively. He is currently a Ph.D. student at Graduate School of Informatics, Kyoto University, Japan. His current interests include web mining and multimedia information retrieval.

Qiang Ma received his Ph.D. degree from Department of Social Informatics, Graduate School of Informatics, Kyoto University in 2004. He was a research fellow (DC2) of JSPS from 2003 to 2004. He joined National Institute of Information and Communications Technology as a research fellow in 2004. From 2006 to 2007, he served as an assistant manager at NEC. From October 2007, he joined Kyoto University and has been an associate professor since August 2010. His general research interests are in the area of databases and information retrieval. His current interests include web mining, multimedia information retrieval and sightseeing informatics.

Yasuhito Asano received his B.S., M.S. and D.S. in Information Science from the University of Tokyo in 1998, 2000 and 2003, respectively. In 2003–2005, he was a research associate of Graduate School of Information Sciences, Tohoku University. In 2006–2007, he was an assistant professor of Department of Information Sciences, Tokyo Denki University. He joined Kyoto University in 2008, and he is currently an associate professor of Graduate School of Informatics. His research interests include Web mining, network algorithms. He is a member of IEEE, DBSJ, and OR Soc. Japan.

Masatoshi Yoshikawa received his the B.E., M.E. and Ph.D. degrees in Information Science from Kyoto University in 1980, 1982 and 1985, respectively. From 1985 to 1993, he was with Kyoto Sangyo University. In 1993, he joined Nara Institute of Science and Technology as an Associate Professor of Graduate School of Information Science. Currently, he is a Professor of Graduate School of Informatics, Kyoto University. His current research interests include XML information retrieval, databases on the Web, and multimedia databases. He is a member of ACM, IPSJ and IEICE.

(Editor in Charge: *Naofumi Yoshida*)