IPSJ Transactions on Databases Vol.5 No.3 117–125 (Sep. 2012)

Regular Paper

Improving Content-based Social Image Retrieval Based on an Image-tag Relationship Model

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

Received: March 20, 2012, Accepted: July 3, 2012

Abstract: With the recent rapid growth of social image hosting websites, such as Flickr, it is easier to construct a large database with social tagged images. We propose an unsupervised approach for automatic ranking social images to improve content-based social image retrieval. We construct an image-tag relationship graph model with both social images and tags. The approach extracts visual and textual information and combines them for ranking by propagating them through the graph links with an optimized mutual reinforcement process. We conduct experiments showing that our approach can successfully use social tags for ranking and improving content-based social image search results, and performs better than other approaches.

Keywords: content based image retrieval, social image, social tag, graph model

1. Introduction

Social image hosting websites, e.g., Flickr [1], have recently become very popular. They are a kind of websites on which users can upload and tag their images for sharing these images with others. This social tagging is similar to keyword annotation in traditional image retrieval systems. An important difference is that keyword annotation requires several experts for annotating images. This requires too much time and labor if the image database is large. Social tagging does not have this problem because a large number of users can participate in tagging task. It is easier to construct a large database with a huge number of tagged images. We denote this kind of image collections with social tags on social image hosting websites as "social image" in our work. The word of "social" in "social image" is to emphasize the characteristics of these huge image collections which have social tags labeled by users.

Social tags have been proven to be effective for providing keyword-based image retrieval and widely used on social image hosting websites. It is regarded that textual information can naturally improve the results of keyword-based image retrieval. However, whether social tags are beneficial for improving content-based image retrieval (CBIR) has not been well investigated in previous work. Services based on CBIR are potential useful applications for social image hosting websites. For example, a user may be interested in other users' images which are similar to his own images; the server can recommend similar images by CBIR to users when they are viewing images. CBIR can provide different image search and recommendation results from keyword-based image retrieval. However, such kinds of services are still not available on social image hosting websites. One of the reasons is that the performance of CBIR still needs to be improved for practical application, though CBIR has a long history and a large amount of research has gone into it [3], [4]. On social image hosting websites, due to large scale of images have been tagged, it will be beneficial for applying CBIR to these websites if social tags can be utilized for improving CBIR results. We therefore focus on improving the performance of CBIR for the social image.

In CBIR, for a query image sample, systems search for content-based similar images from a specific multimedia database by image visual information. Since the query image does not include any textual information, the relationships between the query image and the textual information of other images in the database are hard to be evaluated because of the well-known semantic gap problem. For example, for the query "horse" image in **Fig. 1**, it is hard to know the relationship between the query and the "cat" tag of a "cat" image in the database. The effectiveness of textual information, especially social tags, for improving content-based similar image results is unknown.

On the other hand, social tags are user-generated and folksonomy [2]. In contrast with taxonomy keywords in keyword annotation which uses a number of specific fixed words, social tags have an open vocabulary and are neither exclusive nor hierarchical. This results in social tags having lots of noises. Whether such noisy social tags can be utilized to improve CBIR performance also needs to be investigated.

We observe that in content-based similar image results of a given query image and a given database, relevant images are relatively few while irrelevant images are many. There is a characteristic followed by image semantics that the image semantics of relevant images are alike while the image semantics of irrele-

[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

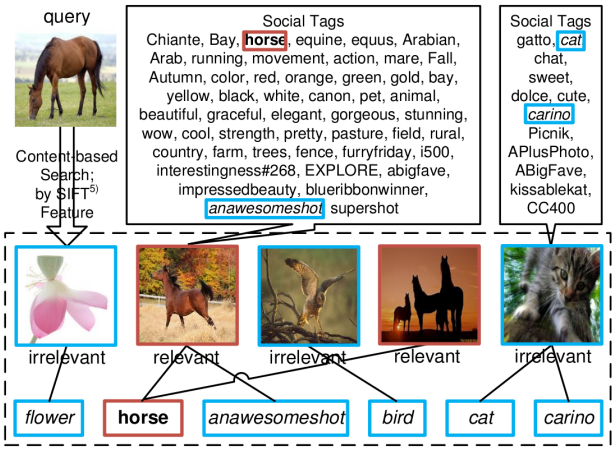


Fig. 1 Query by example, content-based similar image results and social tags.

vant images are diverse. For example, for a query "horse" image, its relevant images have alike "horse" concept while its irrelevant image have diverse concepts such as "cat," "bird," and so on. However in most cases content-based similar image results do not follow this characteristic. Figure 1 shows a query image and its content-based similar images by SIFT feature [5]. These diverse similar images are regarded as "relevant" images in content-based similar image results by CBIR. This is one of the reasons that why the performance of CBIR is unsatisfactory. On the other hand, social tags sometimes follow this characteristic. In Fig. 1, the relevant images have alike tag sets including a "horse" tag, while the irrelevant images have diverse tag sets. It shows that social tags may be able to be used for improving content-based social image search results.

There are existing studies that use visual or textual information for improving keyword-based image retrieval. Because image search results generated from a keyword-based search and a content-based search are different, and social tags have special characteristics which we have introduced, whether the approaches in these studies still have good performance for improving content-based social image retrieval need to be investigated. On the other hand, there are existing studies utilizing both visual and textual information for multimedia information retrieval [6], [7]. They use a linear combination on these two kinds of information. In our work, we try to improve the search performance by fusing the visual and textual information and considering their relationships rather than by linearly combining them.

We propose an unsupervised approach which automatically ranks the images in content-based similar image results. We construct an image-tag relationship graph model, whose vertices are images and their tags, and edges represent image similarity, tag co-occurrence and image-tag annotation relationships. The approach propagates visual and textual information on this graph with a mutual reinforcement process. Figure 1 gives a brief overview of this graph. It shows some of the content-based similar images and social tags on the graph. In the mutual reinforcement ranking process, the good tags (in red and bold) of relevant images contribute more scores on these graph; the bad tags of relevant images, and the good and bad tags (in blue and italic) of irrelevant images contribute less scores on the graph; the irrelevant images contribute less to their tags, while the relevant images contribute more. In other words, a high-ranked image is one to which many high-ranked tags point; a high-ranked tag is a tag that points to many high-ranked images. After several iterations, the relevant images can obtain higher rank scores.

The contributions of this paper are as follows.

- We propose an approach which can utilize social tags to improve content-based image retrieval on social image hosting websites effectively. To the best of our knowledge, it has not been well investigated in previous work. We design an optimized mutual reinforcement process for ranking social images with tags, which outperform a naive mutual reinforcement method.
- We construct a general image-tag relationship graph model to analyze the relationships between social images and tags. We propose an approach which extracts and mutually propagates textual and visual information through the graph links. Experimental results shows that the approach performs better than the other approaches compared in the experimental section.
- We fuse textual and visual information by iteratively propagating the information on the graph. It outperforms existing approaches using linear combination on these two kinds of information.

The remainder of this paper is organized as follows. In Section 2 we give a brief review of related work. In Section 3, we propose our social image ranking approach. In Section 4, we report and discuss the experimental results, and present a summary and discuss future work in Section 5.

2. Related Work

There have been studies related to image ranking for keyword-based image retrieval in unsupervised scenarios. Lin et al. [8] proposed an approach based on textual information only. They proposed a probabilistic relevance model for evaluating the relevance of html document linking to images. Several approaches based on visual information only for improving keyword-based image retrieval have also been proposed. Zitouni et al. [9] presented the similarities of all images in a graph structure, and found the densest component that corresponds to the largest subset of the most similar images. The approach proposed by Zhou et al. [10] performs latent semantic analysis with the visual words of images. The well-known VisualRank approach proposed by Jing et al. [11] applies a random walk method for ranking images. Park et al. [12] and Hsu et al. [13] used clustering methods to adjust the rank results with the distance of a cluster from a query. In contrast, we concentrate on image ranking with social tags for improving content-based image retrieval. To the best of our knowledge, it has not been well investigated in previous work.

Besides the approach using textual information or visual information only, in some areas, some approaches utilize both textual and visual information have been proposed. For example, in early year, Cascia et al. [6] combined visual and textual statistics in a single index vector for content-based search of a WWW image database. Tag Ranking [7] is one of state-of-the-art work which is proposed for ranking the existing social tags of a given image.

It computes linear combination of visual and textual information, and uses it to rank the tags on a tag complete graph with random walk method. Our approach provides a novel way of aggregating textual and visual information based on an image-tag graph model. The experimental results validate the performance of our approach.

On the other hand, user relevance feedback (RF) has a long history and has been widely used in image ranking in supervised scenarios [14]. In early work, approaches have been proposed to adjust the weights of different components of the queries or change the query representation to suit the user's information need. Porkaew et al. [15], [16] proposed query reweighting, query reformulation, query modification, query point movement and query expansion approaches. All these approaches focus on the queries on feature space or the relationships among different features. Many approaches use RF instances as training sets and include a learning process to classify image search results into relevant and irrelevant images. For example, Zhang et al. [17] and Chen et al. [18] proposed approaches using support vector machines (SVM). These approaches always include an offline learning process that uses many queries and corresponding RF instances for learning a query-independent ranking model.

The approaches based on user relevance feedback are proposed for supervised scenarios and need user interactions. In contrast, we concentrate on image ranking in unsupervised scenario without user interactions.

3. Social Image Ranking

We introduce our social image ranking approach in this section. We first define the terms and describe the image-tag relationship model, and then introduce how we extract visual and textual information for analyzing the relationships of images and tags. After that we propose our automatic social image ranking approach in detail.

Our ranking task are formulated as follows. For a given query image q, the content-based image retrieval system returns the top-n content-based similar image results $\mathcal{A} = \{a_1, \cdots, a_n\}$ from a specific multimedia database $\mathcal{D}$. Let s_{iq} be the similarity between q and a_i. We regard $\mathcal{A}$ as the candidate image set and the social tags of images in $\mathcal{A}$ as the candidate tag set $\mathcal{T} = \{t_1, \cdots, t_m\}$. We define $\mathcal{T}_{a_i}$ as the tag set of each image $a_i \in \mathcal{A}$. Our task is to rank the image set $\mathcal{A}$ with the tag set $\mathcal{T}$.

Our approach automatically gathers and aggregates visual and textual information to rank the content-based similar image results. We analyze the relationships among the images and social tags to construct our image-tag relationship model. We propose an approach with an optimized mutual reinforcement process base on the characteristics of this graph model.

3.1 Image-tag Relationship Model

To leverage social image visual information as well as social tag textual information for ranking, we construct a graph model in **Fig. 2** with candidate set $\mathcal{A}$ and $\mathcal{T}$ for analyzing the image-tag relationships. The vertices of the graph model denote social images which represent visual information and their tags which represent textual information. Note that query image q has no textual information.

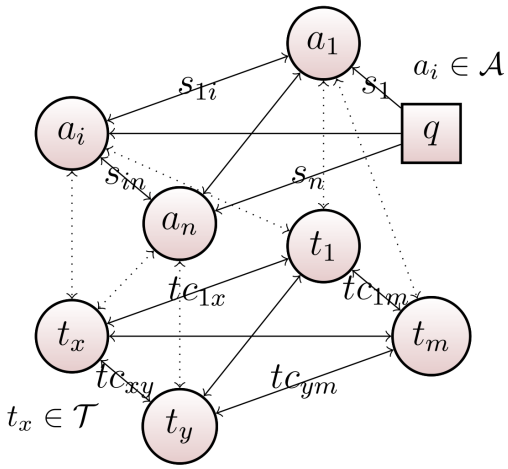


Fig. 2 Image-tag relationship model.

The edges of the graph model denote the relationships among images and tags. There are three kinds of image-tag relationships: image-to-image relationship based on image similarity, tag-to-tag relationship based on tag co-occurrence to images, and image-to-tag annotation relationship. The first two kinds of relationships reflect the intra relationships among images or tags. The third one reflects the inter relationship between images and tags.

3.2 Visual Descriptor

To make use of visual and textual information in our approach, we convert them into visual and textual descriptors. The visual descriptors are based on image similarity. To compute image similarity, we use the following six types of low level features [20]: 64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments and 500-D bag of words based on SIFT.

The distance between image a_i and a_j on low level feature k is computed using the Pearson correlation distance $d(\mathcal{H}_{ik}, \mathcal{H}_{jk})$ defined as

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

$$d(\mathcal{H}_{ik}, \mathcal{H}_{jk}) = \frac{\sum_x (\mathcal{H}'_{ik}(x) * \mathcal{H}'_{jk}(x))}{\sqrt{(\sum_y \mathcal{H}'_{ik}(y)^2) * (\sum_y \mathcal{H}'_{jk}(y)^2)}},$$

where $\mathcal{H}_{ik}$ and $\mathcal{H}_{jk}$ are feature vectors. $\mathcal{N}_k$ is the size of the feature vector k. The image similarity between two images based on multiple features is computed using a weighted sum.

$$s_{ij} = s(a_i, a_j) = \frac{\sum_k w_k d(\mathcal{H}_{ik}, \mathcal{H}_{jk})}{\sum_k w_k}$$

We use $w_k = 1$ for any k in our work. It means that all low level features have same weights. This strategy has usually been used in existing work such as Ref. [21]. For each image a_i in candidate image set $\mathcal{A}$, we propose several optional visual descriptors $vd_i^{(\cdot)}$ defined as

$$vd_i^{(1)} = s_{iq}, \quad vd_i^{(2)} = \sum_j s_{ij}, \quad vd_i^{(3)} = \sum_j exp(-\frac{s_{ij}^2}{2\sigma^2}),$$

where σ in $vd_i^{(3)}$ is the median value of all s_{ij} in $\mathcal{A}$.

3.3 Textual Descriptor

To leverage social tags information, some existing work in other topics use some paired tag co-occurrence measures tc_{xy} for each pair of tags t_x and t_y. For example, in the tag recommendation approach [22], Sigurbjornasson et al. referred two measures. One is a asymmetric measure which is identical to $tc^{(1)}$ in this section. The other is a symmetric measure which is identical to $tc^{(2)}$ in this section. In addition, we propose a symmetric-asymmetric measure $tc^{(3)}$. On the other hand, we also propose a tag importance measure $tc^{(4)}$ for each tag t_x. This measure is not a tag co-occurrence measure. It considers the local tag frequency in candidate tag set $\mathcal{T}$ as well as the global tag frequency in database $\mathcal{D}$ and can evaluate how important the tag t_x is to candidate tag set $\mathcal{T}$ in database $\mathcal{D}$. The definitions of $tc^{(\cdot)}$ are

$$tc^{(1)}_{xy} = \frac{|t_x \cap t_y|_{\mathcal{T}}}{|t_y|_{\mathcal{T}}}, \quad tc^{(2)}_{xy} = \frac{|t_x \cap t_y|_{\mathcal{T}}}{|t_x \cup t_y|_{\mathcal{T}}},$$

$$tc^{(3)}_{xy} = \frac{|t_x \cap t_y|_{\mathcal{T}}}{|t_x|_{\mathcal{T}}} + \frac{|t_x \cap t_y|_{\mathcal{T}}}{|t_y|_{\mathcal{T}}}, \quad tc^{(4)}_{x} = \frac{|t_x|_{\mathcal{T}}}{|t_x|_{\mathcal{D}}}.$$

Here, $|t_x|_{\mathcal{T}}$ means the number of images in candidate tag set $\mathcal{T}$ that contain t_x, $|t_x|_{\mathcal{D}}$ means the number of images in database $\mathcal{D}$ that contain t_x, $|t_x \cap t_y|_{\mathcal{T}}$ means the number of the images that contain both of t_x and t_y, and $|t_x \cup t_y|_{\mathcal{T}}$ means the number of the images that contain t_x or t_y.

For each tag t_x in candidate tag set T, we propose several optional textual descriptors $td^{(\cdot)}$ which are computed by $tc^{(\cdot)}$. They are defined as

$$td^{(1)}_x = \sum_{t_y \in \mathcal{T}} tc^{(\cdot)}_{xy}, \quad td^{(2)}_x = \sum_{t_y \in \mathcal{T}} exp(-\frac{tc^{(\cdot)2}_{xy}}{2\sigma^2}),$$

$$td^{(3)}_x = \begin{cases} tc^{(4)}_x, & if\ |t_x|_{\mathcal{T}} > \delta, \\ 0, & if\ |t_x|_{\mathcal{T}} \le \delta. \end{cases}$$

σ is the mean value of tc_{xy} of which t_x and t_y are in T. δ is a local frequency threshold for ignoring the noisy tags which have low frequency in $\mathcal{T}$ as well as in $\mathcal{D}$ and therefore have high value on tc_x. Note that $td^{(3)}_x$ can be computed by $tc^{(4)}_x$ only.

3.4 Social Image Ranking

We initialize the rank scores Q of images in $\mathcal{A}$ and tags in $\mathcal{T}$ with normalized visual and textual descriptors. Following the image-tag annotation relationships in the graph model, we propagate these rank scores along the links between images and tags. We observe a phenomenon that for an image a_i, when propagating the rank scores from images to tags, if a_i has a high rank score, its related tags will obtain higher rank scores; when propagating the rank scores from tags to images, if the related tags of a_i have high rank scores, a_i will obtain a higher rank score. On the other hand, a similar phenomenon also occurs for a tag t_x. Therefore, we naturally come to the following mutual reinforcement assumption: a high-ranked image for q is the one to which many high-ranked tags point; a high-ranked tag for q is a tag that points to many high-ranked images. The iterative formulas for computing the rank scores are defined as follows.

Initialization: $Q'_0(a_i) = \Phi(vd_i),\ Q'_0(t_x) = \Phi(td_x)$;

Iteration:

$$\begin{cases} Q_{k+1}(t_x) = \alpha\Phi(td_x) + (1-\alpha)\sum_{\vee a_i : t_x \in T_{a_i}} \Phi(vd_i) Q'_k(a_i) \\ Q_{k+1}(a_i) = \beta\Phi(vd_i) + (1-\beta)\sum_{\vee t_x : t_x \in T_{a_i}} \Phi(td_x) Q'_k(t_x) \\ Q'_{k+1}(a_i) = \Phi(Q_{k+1}(a_i)),\ Q'_{k+1}(t_x) = \Phi(Q_{k+1}(t_x)) \end{cases}$$

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

$$\Phi^{(1)}(Q_k(t_x)) = \frac{Q_k(t_x)}{\sqrt{\sum_y Q_k(t_y)^2}},$$

$$\Phi^{(2)}(Q_k(t_x)) = \frac{Q_k(t_x) - min_y\{Q_k(t_y)\}}{max_y\{Q_k(t_y)\} - min_y\{Q_k(t_y)\}}.$$

The iteration parameters α and β are damping factors. k is the number of iteration steps. $Q'_k(\cdot)$ is the normalized rank score of $Q_k(\cdot)$. We propose two alternative normalization functions $\Phi(\cdot)$ here.

Content-based image similarity to the query 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 descriptors and textual descriptors as the weights of images and tags in the iterations. These weights represent the importance of these images and tags on the graph. As demonstrated in Section 4.3, the weighting factors in the iterative formulas play an important role in performance enhancement.

4. Experiment

4.1 Experimental Settings

The dataset we use for experiment is NUS-WIDE [20]. It is created by downloading images and their social tags from social image hosting website Flickr [1]. It has 269,648 images and about 425,000 unique original tags. For images, it provides six types of low-level features extracted from the images, which we have introduced in Section 3.2. For tags, the authors of this dataset set several rules to filter the original tag set. They delete the tags with too low frequency. The low frequency threshold is set to 100. They also remove the tags that does not exist in WordNet. At the end, they provide 5,018 unique tags. We keep this filtering in our experiment for the following reasons. It reduces the noises in the tag set. It also reduces the size of candidate tag set $\mathcal{T}$ and the number of links between images and tags, which can reduce the time cost in the ranking computation.

NUS-WIDE also provides image annotation ground-truth of 81 concepts for the entire dataset, but it does not appoint a query sample set and provide ground-truth for content-based image retrieval. We need to construct them by ourselves for our experiment. In our experiment, we randomly choose 100 images as a query image set from the entire dataset for our evaluation. We choose 20 queries in this query image set as a training set for parameters tuning and the other 80 queries as a testing set for performance evaluation. Note that there is no textual information available for these queries. Jain et al. [23] claim that such size of query image set is comparable for experiments in image ranking area. For each query, we rank the images in top-n content-based similar image results, where the cut-off size $n = 100$. The images in content-based similar image results are labeled with

Table 1 Sample of relevance levels for NDCG.

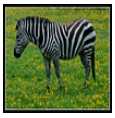
Query	Image Samples				
					
Relevance Level	4	3	2	1	0

Table 2 Parameter selections.

Parameters				NDCG@100
vd	tc	td	Φ	
$\mathbf{vd^{(1)}}$	$tc^{(4)}$	$td^{(3)}$	$\Phi^{(2)}$	**0.6761**
$vd^{(2)}$	$tc^{(4)}$	$td^{(3)}$	$\Phi^{(2)}$	0.6445
$vd^{(3)}$	$tc^{(4)}$	$td^{(3)}$	$\Phi^{(2)}$	0.6314
$vd^{(1)}$	$tc^{(1)}$	$td^{(1)}$	$\Phi^{(2)}$	0.5628
$vd^{(1)}$	$tc^{(2)}$	$td^{(1)}$	$\Phi^{(2)}$	0.5201
$vd^{(1)}$	$tc^{(3)}$	$td^{(1)}$	$\Phi^{(2)}$	0.5496
$vd^{(1)}$	$tc^{(1)}$	$td^{(2)}$	$\Phi^{(2)}$	0.53423
$vd^{(1)}$	$tc^{(2)}$	$td^{(2)}$	$\Phi^{(2)}$	0.53078
$vd^{(1)}$	$tc^{(3)}$	$td^{(2)}$	$\Phi^{(2)}$	0.53421
$vd^{(1)}$	$\mathbf{tc^{(4)}}$	$\mathbf{td^{(3)}}$	$\Phi^{(2)}$	**0.6761**
$vd^{(1)}$	$tc^{(4)}$	$td^{(3)}$	$\Phi^{(1)}$	0.6661
$vd^{(1)}$	$tc^{(4)}$	$td^{(3)}$	$\mathbf{\Phi^{(2)}}$	**0.6761**

five relevance levels by us, according to their visual and semantic relevance to the query images. The range of relevance levels is from 0 to 4: irrelevant (0), weakly relevant (1), partially relevant (2), relevant (3), and very relevant (4). **Table 1** shows an intuitive example of different relevance levels. The evaluation metric used in our experiment is Normalized Discounted Cumulative Gain (NDCG) [19]. NDCG is an effective metric often used in information retrieval for evaluating the rank results with relevance levels. For a given result, when more images with higher relevance scores are ranked higher, the NDCG score of this result is higher. It is defined as follows,

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

$r(j)$ is the relevance level of the image at rank j. Z_k is a normalization constant and equal to the maximum DCG value that the top-k ranked images can reach, so that NDCG score is equal to 1 for the optimal results of which the relevance scores have a descending order. We evaluate the performance with the average NDCG value of query images.

4.2 Parameters Selection

We use the training set with 20 queries for parameters selection. The upper limit of iteration times of our approach is set to 10. To select proper visual descriptor $vd^{(\cdot)}$, we observe the results of different $vd^{(\cdot)}$ on NDCG@100 metric by keeping all of other parameters unchanged. **Table 2** illustrates the results. This table includes three blocks. The first block fixes tc, td and Φ, and changes vd; The second block fixes vd and Φ, and changes tc and td; The third block fixes tc, td and vd, and changes Φ. Note that the NDCG@100 of the content-based similar images results is 0.5911. We select $vd^{(1)}$ as the visual descriptor in our approach because it has better performance than the other two descriptors in the experiment. The selection of textual descriptor $td^{(\cdot)}$, tag co-occurrence $tc^{(\cdot)}$ and normalization function $\Phi^{(\cdot)}$ are similar. We select $\Phi^{(2)}$, $tc^{(4)}$, $td^{(3)}$ in our approach. $td_x^{(2)}$ which is

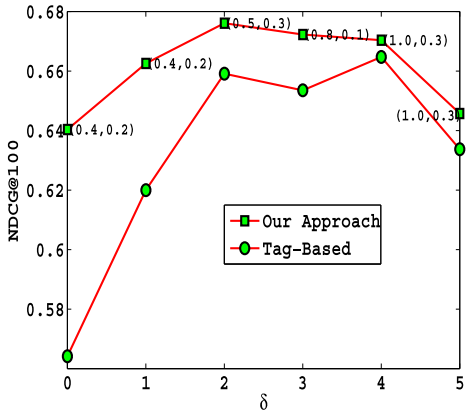


Fig. 3 δ of $td_x^{(3)}$, $\arg\max_{(\alpha,\beta)}(NDCG)$.

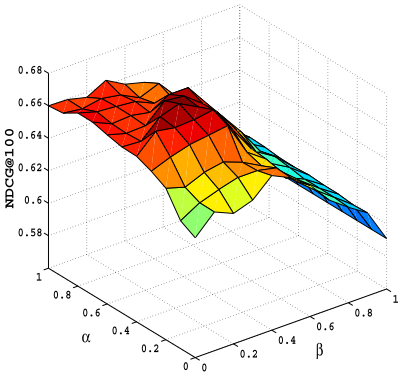


Fig. 4 Parameters α and β, $\delta = 2$.

based on Gaussian related similarity does not have better performance here because social tags are not densely distributed around content topics and the average tag frequency is low.

The proper local frequency threshold δ of $td^{(3)}$ is different for different approaches. **Figure 3** shows the maximal NDCG@100 value that our approach can reach with different δ. In this figure, e.g., if $\delta = 2$, our approach can reach the maximal NDCG@100 value when (α,β) is set to (0.5,0.3). **Figure 4** shows how we set and select proper iteration parameters α and β in our mutual reinforcement approach, we choose their candidate values by an interval of 0.1 in the range of $[0, 1]$ and obtain 121 pairs of candidate values. We run our approach with these pairs on the training set and observe the performance on the NDCG@100 metric. According to Fig. 3 and Fig. 4, we select $\delta = 2$ and (α,β) as $(0.5, 0.3)$ for our approach.

Note that when $\beta = 1$, it means the iteration formula of an image rank score has degenerated into depending on the visual descriptor only. Because of using $vd^{(1)}$ as the visual descriptor, the ranking result is equal to the content-based similar image results. Figure 4 also shows that for any (α,β) in the range, our approach performs not worse than content-based similar image

results.

4.3 Experimental Results

We compare our approach with four other approaches as well as with the content-based similar image results. The first one is based on a typical and well-known approach, VisualRank [11], which is a visual-based approach and utilizes visual information only. The second one is a tag-based approach that uses social tags but without mutual reinforcement process. It is an approach which utilizes textual information only. The third one is a naive mutual reinforcement approach which refers to the well-known HITS [25] approach. The forth one refers to Tag Ranking approach [7] and combines both visual and textual information for ranking. The parameters selection and tuning is carried out for all these approaches. We compare the best results these approaches can generate.

Visual-based Approach (VisualRank): This approach does not use any social tag information. It applies PageRank [24] approach and uses a random walk method on the image complete graph in which vertices are the candidate images, and uses content-based image similarity for computing the transition matrix. The iteration formula is as follows.

$$Q_{k+1}(a_i) = (1-\gamma) * \frac{1}{n} + \gamma * \sum_j (Q_k(a_j) * \frac{s_{ij}}{\sum_x s_{xj}}),$$

$$Q_0(a_i) = \Phi^{(2)}(vd_i^{(1)}).$$

We follow the settings in VisualRank and set damping factor γ to 0.85. n is the size of the candidate image set $\mathcal{A}$. We want to confirm that social tags are beneficial for ranking content-based similar image results and show that our approach performs better than VisualRank for our topic.

Tag-based Approach: To show that our mutual reinforcement process can use social tag information more effectively, we design a tag-based approach that uses social tag information but without a mutual reinforcement process. We compute the rank score of candidate image a_i by using the following formula.

$$Q(a_i) = \sum_{\vee a_i : t_x \in T_{a_i}} \Phi^{(2)}(td_x^{(3)})$$

We use the same text descriptor $td_x^{(3)}$ with our approach. According to Fig. 3, we set $\delta = 4$. We also evaluate its performance setting $\delta = 0$, and compare with our approach setting $\delta = 0$. Furthermore, note that it is not a pure textual-only-based approach in our scenario because the candidate image and tag set are generated by visual information.

HITS: The mutual reinforcement process is well-known and some approaches based on it have been proposed in other areas, e.g., HITS approach [25] for rating web pages. We set this naive approach to show that the introduction of the weights in the 2nd term of the iterative formulas, and the cautious selections on the visual and textual descriptors and normalization function, yield better performance than the naive one. In this naive approach, we choose $vd^{(1)}$, $\Phi^{(1)}$, $tc^{(2)}$ and $td^{(1)}$ as the parameters. (α, β) is chosen as $(0.0, 0.9)$ with which this naive approach can reach maximal NDCG@100 value on the training set. Note that it is not too naive because it still uses a good visual descriptor and damping factors. The rule of parameters chosen here is to choose some intuitive parameters.

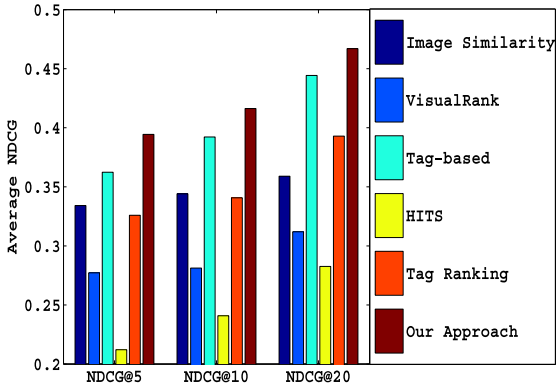


Fig. 5 Comparison of results.

Initialization: $Q'_0(a_i) = \Phi^{(1)}(vd_i^{(1)})$, $Q'_0(t_x) = \Phi^{(1)}(td_x^{(1)})$;

Iteration:

$$\begin{cases} Q_{k+1}(t_x) = \alpha\Phi^{(1)}(td_x^{(1)}) + (1-\alpha)\sum_{\vee a_i : t_x \in T_{a_i}} Q'_k(a_i) \\ Q_{k+1}(a_i) = \beta\Phi^{(1)}(vd_i^{(1)}) + (1-\beta)\sum_{\vee t_x : t_x \in T_{a_i}} Q'_k(t_x) \\ Q'_{k+1}(t_x) = \Phi^{(1)}(Q_{k+1}(t_x)), Q'_{k+1}(a_i) = \Phi^{(1)}(Q_{k+1}(a_i)) \end{cases}$$

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

Tag Ranking: We refer the approaches proposed in Ref. [7] which ranks the existing tags of a given image. It computes linear combination of visual and textual information, and uses it to rank the tags on a tag complete graph with random walk method. Tag Ranking approach can not be utilized directly for our topic. Even the inverted scenario of Tag Ranking is to rank the images to which a given tag is labeled, which is also different from our topic. In our experiment, this baseline approach refers the method that Tag Ranking used for visual and textual information combination and ranking.

We use the testing set with 80 queries for performance evaluation. **Figure 5** illustrates the evaluation of NDCG@5, NDCG@10 and NDCG@20 metrics on ranking the top-100 images in content-based similar image results. It is equivalent to the evaluation of NDCG on top-5%, 10% and 20% images in our ranking results. In contrast with content-based similar image results, all metrics are improved with our approach. Our approach performs the best among all of the approaches here.

4.4 Discussion

Our approach performs better than the visual-based approach in our content-based social image ranking scenario. Although VisualRank performs well in Jing et al.'s work [11], the initial image results in that work are from keyword-based image retrieval, and the ranking is based on image similarity. In other words, it uses both visual and textual information to generate the final results. But in our work, since the initial image results are content-based, the visual-based approach can only use visual information. It illustrates that using social tag information for ranking the content-based image search results is beneficial. Furthermore our approach also has better time complexity than VisualRank because our approach computes less links on the graph in real time

Table 4 Social image ranking examples.

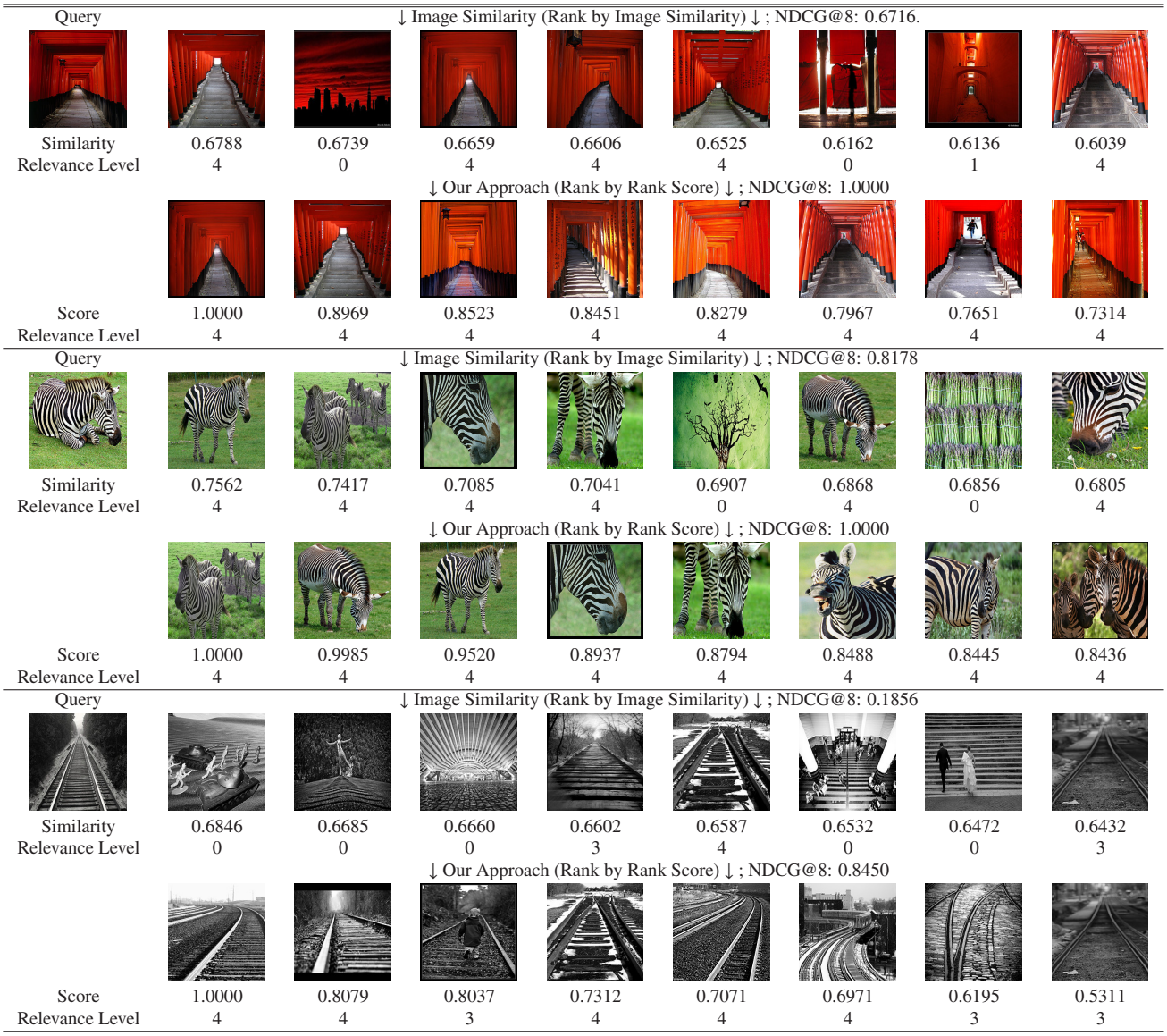


Table 3 Number of images of which all tags have $td_x^{(3)} = 0$.

δ	0	1	2	3	4	5
Number	30	552	1175	1770	2286	2696

iterations. In our experiments, on the average running time for all queries, VisualRank costs 0.076 seconds while ours costs 0.031 seconds.

In contrast with the tag-based approach which does not use our mutual reinforcement process, our approach which is also based on textual information performs better. The tag-based approach also performs better than VisualRank. It shows that social tags are beneficial for ranking content-based similar image results and our approach can use them more effectively.

On the other hand, among all of the approaches for comparison, tag-based approach has the most approximate performance to our approach. However, our approach performs better than tag-based approach not only on the quality of search results, but also on the property of noise resistance. As proposed in Section 3.3, parameter δ of textual descriptor $td_x^{(3)}$ is a local frequency threshold for ignoring noisy tags. **Table 3** shows the number of images

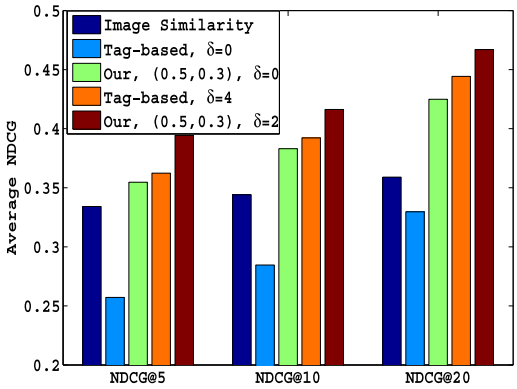


Fig. 6 Noise resistance.

of which all tags have $td_x^{(3)} = 0$. When this number is too small, it means there are lots of noisy tags included in the computation for ranking; when this number is too large, it means there are some good tags removed from the computation of ranking. The experimental data in Fig. 3 follows this statement. **Figure 6** and Table 3 show that when $\delta = 0$, which means noisy tags are numerous,

the performance of tag-based decreases a lot and is lower than the content-based social image search results; the performance of our approach decreases a little and can still improve the content-based social image search results effectively.

Tag Ranking performs better than VisualRank in our scenario. Because both of them utilize random walk method, while Tag Ranking uses textual information and VisualRank does not, it shows that utilizing social tags can improve content-based social image search results. On the other hand, our approach performs better than both of these two random walk based approaches. It shows that our optimized mutual reinforcement method on image-tag relationship graph, which is a multi-media graph, performs better than random walk based approaches on image complete graph, which is a single-media graph, on ranking social images in our scenario.

From the aspect of aggregating visual and textual information, Tag Ranking utilizes a linear combination method; our approach proposes a mutual propagation process to fuse visual and textual information. Our approach performs better than Tag Ranking. It shows that our approach can aggregate visual and textual information for ranking more effectively in our topic.

Our approach performs better than HITS, a naive mutual reinforcement approach which outperforms VisualRank. It shows that a mutual reinforcement process is useful in our ranking scenario, but a naive mutual reinforcement approach without an optimized design still can not generate better rank results than the content-based similar image results from a statistical viewpoint. Our proposed mutual reinforcement approach can improve the content-based similar image results effectively.

Table 4 illustrates some social image ranking samples. Our approach can promote the rankings of relevant images and demote the rankings of irrelevant images effectively.

5. Conclusion

In this paper, we confirm that we can successfully use social tags to improve content-based social image search results. We propose an approach with a mutual reinforcement process fusing both visual and textual information on an image-tag relationship graph model. The experiments illustrate that our approach can reach the goals and performs better than the other approaches.

For future work, we will extend our work to keyword-based social image retrieval. We plan to construct image and tag candidate sets with keyword-based image search results, apply and evaluate our approach.

Acknowledgments This work was supported by MEXT/ JSPS KAKENHI Grant Number (20300036).

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] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A. and Jain, R.: Content-based image retrieval at the end of the early years, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No.12, pp.1349–1380 (2002).
[4] Datta, R., Joshi, D., Li, J. and Wang, J.Z.: Image Retrieval: Ideas, Influence, and trends of the new age, *ACM Computing Surveys*, Vol.40, Iss.2, No.5 (2008).
[5] Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol.60, No.2, pp.91–110 (2004).
[6] Cascia, M.L., Sethi, S. and Sclaroff, S.: Combining textual and visual cues for content-based image retrieval on the world wide web, *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp.24–28 (1998).
[7] 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).
[8] Lin, W.H., Jin, R. and Hauptmann, A.: Web Image Retrieval Re-Ranking with Relevance Model, *Proc. 2003 IEEE/WIC International Conference on Web Intelligence* (*WI'03*), pp.242–248 (2003).
[9] Zitouni, H., Sevil, S., Ozkan, D. and Duygulu, P.: Re-ranking of web image search results using a graph algorithm, *Proc. 19th International Conference on Pattern Recognition* (*ICPR'08*), pp.1–4 (2008).
[10] Zhou, W.G., Tian, Q., Yang, L.J., Li, H.Q.: Latent visual context analysis for image re-ranking, *Proc. ACM International Conference on Image and Video Retrieval* (*CIVR'10*), pp.205–212 (2010).
[11] Jing, Y. and Baluja, S.: VisualRank: Applying PageRank to Large-Scale Image Search, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.30, No.11, pp.1877–1890 (2008).
[12] Park, G., Baek, Y. and Lee, H.: Re-ranking algorithm using post-retrieval clustering for content-based image retrieval, *Information Processing and Management*, Vol.41, No.2 (2005).
[13] Hsu, W., Kennedy, L. and Chang, S.: Video search reranking via information bottleneck principle, *Proc. 14th Annual ACM International Conference on Multimedia* (*MULTIMEDIA '06*), pp.35–44 (2006).
[14] Yong, R., Huang, T.S., Ortega, M. and Mehrotra, S.: Relevance feedback: a power tool for interactive content-based image retrieval, *IEEE Trans. Circuits and Systems for Video Technology*, Vol.8, Iss.5, pp.644–655 (1998).
[15] Porkaew, K., Mehrotra, S. and Ortega, M.: Query reformulation for content based multimedia retrieval in MARS, *Proc. IEEE International Conference on Multimedia Computing and Systems* (*ICMCS'99*), pp.747–751 (1999).
[16] Porkaew, K., Chakrabarti, K. and Mehrotra, S.: Query Refinement for Multimedia Similarity Retrieval in MARS, *Proc. 7th ACM International Conference on Multimedia* (*Multimedia'99*), pp.235–238 (1999).
[17] Zhang., L., Lin, F.Z. and Zhang, B.: Support vector machine learning for image retrieval, *Proc. 2001 International Conference on Image Processing* (*ICIP'01*), pp.721–724 (2001).
[18] Chen, Y.Q., Zhou, X.S. and Huang, T.S.: One-class SVM for learning in image retrieval, *Proc. 2001 International Conference on Image Processing* (*ICIP'01*), pp.34–37 (2001).
[19] Jarvelin, K. and Kekalainen, J.: Cumulated gain-based evaluation of IR techniques, *ACM Trans. Information Systems* (*TOIS*), Vol.20, Iss.4 (2001).
[20] Chua, T.S., Tang, J.H., Hong, R.C., Li, H.J., Luo, Z.P. and Zheng, Y.T.: NUS-WIDE: A Real-World Web Image Database from National University of Singapore, *Proc. ACM International Conference on Image and Video Retrieval* (CIVR'09) (2009).
[21] Van de Sande, K.E.A., Gevers, T. and Snoek C.G.M.: Evaluating Color Descriptors for Object and Scene Recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.32, Iss.9, pp.1582–1596 (2010).
[22] 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).
[23] Jain, V. and Varma, M.: Learning to re-rank: Query-dependent image re-ranking using click data, *Proc. 20th International Conference on World Wide Web* (*WWW'11*), pp.277–286 (2011).
[24] 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).
[25] Kleinberg, J.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol.46, No.5, pp.604–632 (1999).

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 multimedia database, Web mining, and multimedia information retrieval.

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 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: *Chiemi Watanabe*)