Tag Ranking*

Dong Liu
School of Computer Sci.& Tec.
Harbin Institute of Technology
Harbin, 150001, P.R.China
dongliu@hit.edu.cn

Xian-Sheng Hua
Microsoft Research Asia
49 Zhichun Road
Beijing, 100190, P.R.China
xshua@microsoft.com

Linjun Yang
Microsoft Research Asia
49 Zhichun Road
Beijing, 100190, P.R.China
linjuny@microsoft.com

Meng Wang
Microsoft Research Asia
49 Zhichun Road
Beijing, 100190, P.R.China
mengwang@microsoft.com

Hong-Jiang Zhang
Microsoft Adv. Tech. Center
49 Zhichun Road
Beijing, 100190, P.R.China
hjzhang@microsoft.com

ABSTRACT

Social media sharing web sites like Flickr allow users to annotate images with free tags, which significantly facilitate Web image search and organization. However, the tags associated with an image generally are in a random order without any importance or relevance information, which limits the effectiveness of these tags in search and other applications. In this paper, we propose a tag ranking scheme, aiming to automatically rank the tags associated with a given image according to their relevance to the image content. We first estimate initial relevance scores for the tags based on probability density estimation, and then perform a random walk over a tag similarity graph to refine the relevance scores. Experimental results on a 50,000 Flickr photo collection show that the proposed tag ranking method is both effective and efficient. We also apply tag ranking into three applications: (1) tag-based image search, (2) tag recommendation, and (3) group recommendation, which demonstrates that the proposed tag ranking approach really boosts the performances of social-tagging related applications.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

Keywords

Flickr, tag ranking, random walk, recommendation, search

1. INTRODUCTION

Recent years have witnessed an explosion of community-contributed multimedia content available online (e.g. Flickr, Youtube, and Zooomr). Such social media repositories allow users to upload personal media data and annotate content with descriptive keywords called tags. With the rich tags as metadata, users can more conveniently organize and access shared media content.

Figure 1: An exemplary image from Flickr and its associated tag list. There are many imprecise and meaningless tags in the list and the most relevant tag "dog" is not at the top positions.

We take Flickr [1], one of the earliest and most popular social media sharing web sites, as an example to study the characteristics of these user-created tags. As pointed out in [2] , the principal purpose of tagging is to make Flickr photos better accessible to the public. However, existing studies reveal that many tags provided by Flickr users are imprecise and there are only around 50% tags actually related to the image [3]. Furthermore, the importance or relevance levels of the tags cannot be distinguished from current tag list, where the order is just according to the input sequence and carries little information about the importance or relevance. Fig. 1 is an exemplary image from Flickr, from which we can see that the most relevant (or descriptive) tag is actually "dog", but this cannot be discovered from the tag list directly.

Fig. 2 shows the position (in terms of the tag list) distribution of the most important tags. It is generated from 1,200 random Flickr images with at least 10 tags. For each image, its most relevant tag from the list is labeled based on the majority voting of five volunteers. As can be seen, only less than 10% of the images have their most relevant tag at the top position in their attached tag list. This illustrates that

*This work was performed at Microsoft Research Asia.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

the tags are almost in a random order in terms of relevance to the associated image[1].

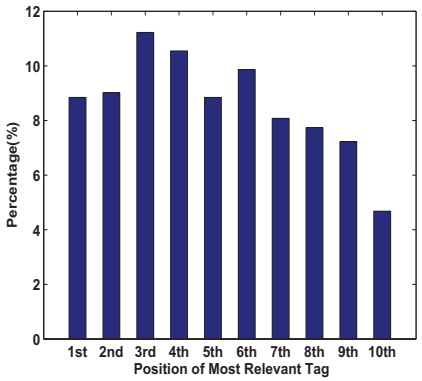


Figure 2: Percentage of images that have their most relevant tag at the n-th position in the associated tag list, where $n = 1, 2, 3 \ldots, 10$.

The lack of relevance information in the tag list has significantly limited the application of tags. For example, in Flickr tag-based image search service[2] , currently it cannot provide the option of ranking the tagged images according to relevance level to the query[3]. However, relevance ranking is important for image search, and all of the popular image search engines, like Google and Live, rank the search results by relevance. Fig. 3 shows the search result for the query "bird" ranking by "interestingness". We can observe that almost half of these images are actually irrelevant or weakly relevant to the query "bird". As can be seen, although "bird" has been tagged to these weakly relevant images, it is not the most relevant one compared with other tags. Meanwhile, for the highly relevant results, "bird" is indeed the most relevant tag. If we can rank the associated tags for each image based on the relevance level, then a better ranking for the tag based image search can be obtained(see Figure 14).

Besides the application of tag ranking into tag-based image search, it can also be used in many of other social-tagging related services, including tag recommendation and Flickr group recommendation.

[1]It is worth mentioning that most of the tags tend to appear in the first 10 positions though they are nearly randomly distributed. For example, a Flickr image has more than 10 tags in average, but according to Fig. 2 we can see that 70% of the images have their most relevant tags at the first 10 positions. This indicates that there is still a trend that top tags are more relevant, although the trend is weak. Actually it is easy to understand that there will be certain correlation between the input sequence and the relevance levels of tags, since this coincides with most users' habit. But the correlation is not strong, and in experiments we will take the original tag lists as baseline to show that our tag ranking approach provides a much better order .

[2]http://www.flickr.com/search/?q=cat&m=tags

[3]Currently Flickr offers two options in the ranking for tag-based image search. One is "most recent", which ranks the most recently uploaded images on the top and the other is "most interesting", which ranks the images by "interestingness", a measure that takes click-through, comments, etc, into account, as stated in http://www.flickr.com/explore/interesting.

Figure 3: Search results of query "bird" in Flickr, which are ordered by "interestingness". For each image, only top tags are displayed due to space limit.

- Image tag recommendation. Tag recommendation is to recommend a set of tags for one image based on existing tags so that users only need to select the relevant tags, instead of to type the tags manually. If the tags have been ranked according to their relevance to the image, a better recommendation can be obtained. For example, for each uploaded image, we can firstly find the K nearest neighbors based on low-level visual features, and then the top ranked tags of the K neighboring images are collected and recommended to the user. We will show that this method can achieve highly satisfactory recommendations, even better than the tags input by Flickr users.

- Flickr group recommendation. Flickr group is a collection of images created by users with certain common interest. If tags are ranked appropriately, for each uploaded image, we use the top tags in its ranked tag list to search for related Flickr groups and recommend to users. Experiments will show this method is able to recommend more suitable groups to users and this process is totally automatic without any user interaction.

Though important, tag ranking, has not been studied in information retrieval and multimedia societies, as we have conducted in webpage/image/video ranking [4, 5]. As to our best knowledge, this is the first study to this problem.

In this paper we propose a tag ranking approach in which the tags of an image can be automatically ranked according to their relevance with the image. To accomplish the

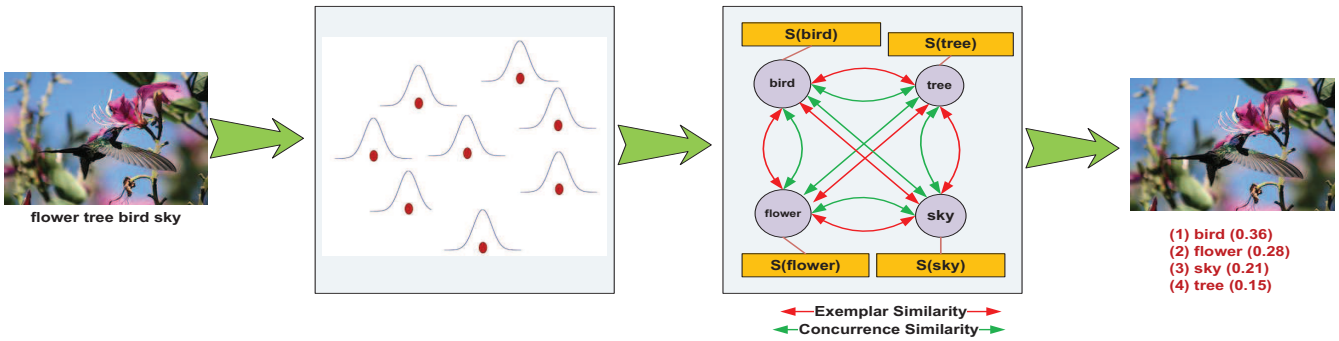


Figure 4: The illustrative scheme of the tag ranking approach. A probabilistic method is first adopted to estimate tag relevance score. Then a random walk-based refinement is performed along the tag graph to further boost tag ranking performance.

ranking, we first adopt a probabilistic approach to estimate the initial relevance score of each tag for one image individually, and then refine the relevance scores by implementing a random walk process over a tag graph in order to mine the correlation of the tags. In the construction of tag graphs, we have combined an exemplar-based approach and a concurrence-based approach to estimate the relationship among tags. The whole process is automatic and do not need any manually labeled training data. Experimental results demonstrate that the proposed scheme is able to rank Flickr image tags according to their relevance levels.

The rest of this paper is organized as follows. We describe the tag ranking scheme in Section 2 and provide empirical justifications in Section 3. In Section 4, we introduce the three application scenarios and the associated experimental results. Then we introduce related work in Section 5. Finally, we conclude the paper in Section 6.

2. TAG RANKING

In this section, we will introduce our tag ranking method. We firstly give an overview of our tag ranking approach, and then introduce the probabilistic relevance score estimation and random walk-based refinement in detail.

2.1 Overview

As illustrated in Fig. 4, the tag ranking scheme mainly consists of two steps: initial probabilistic tag relevance estimation and random walk refinement. Given an image and its associated tags, we first estimate the relevance score of each tag individually through a probabilistic approach. We will simultaneously consider the probability of the tag given the image and the descriptive ability of the tag in the relevance score estimation, and we show that it can be accomplished by using the Kernel Density Estimation (KDE) [15]. Although the scores obtained in this way reflect the tag relevance, the relationships among tags have not been taken into account. Thus we further perform a random walk-based refinement to boost tag ranking performance by exploring the relationship of tags. Finally, the tags of the image can be ranked according to their refined relevance scores. In the next two sub-sections, we will detail the probabilistic relevance score estimation method and the random walk-based refinement process, respectively.

2.2 Probabilistic Tag Relevance Estimation

First, we estimate the relevance scores of the tags from the probabilistic point of view. Given a tag t, its relevance score to an image x is defined as

$$s(t,x) = p(t|x)/p(t) \tag{1}$$

Now we will explain the rationality of Eq. 1. In fact, the most straightforward way is to directly regard $p(t|x)$ as the relevance score, since it indicates the probability of tag t given image x. However, the tag may not be so descriptive when it appears too frequently in the dataset. For example, for the tag "image", the probability $p(t|x)$ will be always 1, but obviously this tag is non-informative. Therefore, we normalize $p(t|x)$ by $p(t)$, i.e., the prior probability of the tag, to penalize frequently-appearing tags. This principle has actually been widely investigated in information retrieval, e.g., in the design of tf-idf features [16].

Based on Bayes' rule, we can easily derive that

$$s(t,x) = \frac{p(x|t)p(t)}{p(x)p(t)} = \frac{p(x|t)}{p(x)} \tag{2}$$

where $p(x)$ and $p(x|t)$ are the prior probability density function and the probability density function of images conditioned on the tag t, respectively. Since the target is to rank the tags for the individual image and $p(x)$ is identical for these tags, we can simply redefine Eq. 2 as

$$s(t,x) \doteq p(x|t) \tag{3}$$

We adopt the classical Kernel Density Estimation (KDE) method to estimate the probability density function $p(x|t)$. Denote by X_i the set of images that contain tag t_i, the KDE approach measures $p(x|t_i)$ as

$$s(t_i,x) = p(x|t_i) = \frac{1}{|X_i|}\sum_{x_k \in X_i} K_\sigma(x - x_k) \tag{4}$$

where $|X_i|$ is the cardinality of X_i and K_σ is the Gaussian kernel function with the radius parameter σ, i.e.,

$$K_\sigma(x - x_k) = exp(-\frac{||x - x_k||^2}{\sigma^2}) \tag{5}$$

The relevance score computed in Eq. 4 actually has a very intuitive explanation. For each image x, the neighbors X_i

can be regarded as its friends. The sum of the similarities estimated based on Gaussian kernel function can be regarded as the soft voting from the friends. Therefore, such relevance is actually estimated based on "collective intelligence" from friend images.

2.3 Random Walk-Based Refinement

The probabilistic tag relevance estimation, which takes the image "friends" into consideration, has not taken into account the relationship among tags which will be helpful. For example, consider the image which has a lot of closely related tags, such as "cat", "animal" and "kitten", and an isolated tag such as "Nikon". In this case, intuitively we can infer that the isolated tag "Nikon" is less descriptive than the others.

To investigate the relationship between tags, we perform random walk over the tag graph for each image to propagate the relevance scores among tags. The nodes of the graph are the tags of the image and the edges are weighted with pairwise tag similarity. Here we propose two tag similarity measurements, i.e., exemplar similarity and concurrence similarity, and combine them in the tag graph construction.

2.3.1 Tag graph construction

We estimate the tag exemplar similarity and concurrence similarity and then combine them to form the weights of tag graph edges.

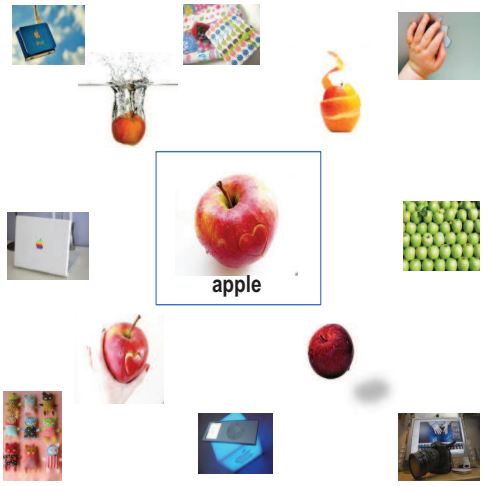


Figure 5: The images containing tag "apple" are diversified and only using nearest neighbors of the central image will help produce better representative exemplars.

We first estimate tag exemplar similarity from visual clue. For a tag t associated with an image x, we collect the N nearest neighbors from the images containing tag t, and these images are regarded as the exemplars of the tag t with respect to x (more exactly, the images are the local exemplars of the tag since we have only used the neighbors of image x). The purpose of adopting nearest neighbor strategy here is to avoid noise introduced by polysemy. Take the central image in Fig. 5 as an example. The images in Fig. 5 are all with tag "apple" and we can see that they are highly diversified. But actually the tag "apple" of the central image only indicates a fruit. So, only using the neighbors of this central image will reduce noise and lead to better tag representative collection.

Denote by Γ_t the representative image collection of tag t. Then the exemplar similarity between tags t_i and t_j is defined as follows

$$\varphi_e(t_i, t_j) = exp(-\frac{1}{N * N}\Sigma_{x\in\Gamma_{t_i}, y\in\Gamma_{t_j}}\frac{||x-y||^2}{\sigma^2}) \quad (6)$$

Note that here we have used the same radius parameter σ as in Eq. 5. The whole process can be illustrated in Fig. 6.

We then define the concurrence similarity between tags based on their co-occurrence. Analogous to the principle of Google distance [17], we first estimate the distance between two tags t_i and t_j as follows.

$$d(t_i, t_j) = \frac{max(logf(t_i), logf(t_j)) - logf(t_i, t_j)}{logG - min(logf(t_i), logf(t_j)))} \quad (7)$$

where $f(t_i)$ and $f(t_j)$ are the numbers of images containing tag t_i and tag t_j respectively and $f(t_i, t_j)$ is the number of images containing both t_i and t_j. These numbers can be obtained by performing search by tag on Flickr website using the tags as keywords, as illustrated in Fig. 7. Moreover, G is the total number of images in Flickr. The concurrence similarity between tag t_i and tag t_j is then defined as

$$\varphi_c(t_i, t_j) = exp(-d(t_i, t_j)) \quad (8)$$

To explore the complementary nature of exemplar similarity and concurrence similarity, we combine them as

$$s_{ij} = s(t_i, t_j) = \lambda \cdot \varphi_e(t_i, t_j) + (1-\lambda) \cdot \varphi_c(t_i, t_j) \quad (9)$$

where λ belongs to $[0, 1]$. In Section 3 we will demonstrate that their combination is better than using each one individually. The combined similarity is used as the weight of the edge between t_i and t_j in the tag graph.

2.3.2 Random walk over tag graph

Random walk methods have been widely applied in machine learning and information retrieval fields [18, 19, 20]. Here we perform random walk process over the tag graph in order to boost the performance of tag ranking by using the relationship among tags. Given a tag graph with n nodes, we use $r_k(i)$ to denote the relevance score of node i at iteration k. Thus, the relevance scores of all the nodes in the graph at iteration k form a column vector $\mathbf{r}_k \equiv [r_k(i)]_{n\times 1}$. Denote by $\mathbf{P}$ an n-by-n transition matrix. Its element p_{ij} indicates the probability of the transition from node i to node j and it is computed as

$$p_{ij} = \frac{s_{ij}}{\sum_k s_{ik}} \quad (10)$$

where s_{ij} denotes the pairwise tag similarity(see Eq. 9) between node i and j.

The random walk process is thus formulated as

$$r_k(j) = \alpha \sum_i r_{k-1}(i)p_{ij} + (1-\alpha)v_j \quad (11)$$

where v_j is the initial probabilistic relevance score of tag t_j, and α is a weight parameter that belongs to $(0, 1)$. The above process will promote the tags that have many close neighbors and weaken isolated tags. Now we prove the convergence of the iteration of Eq. 11.

THEOREM 1. *The iteration of Eq. 11 converges to a fixed point* $\boldsymbol{r}_\pi$.

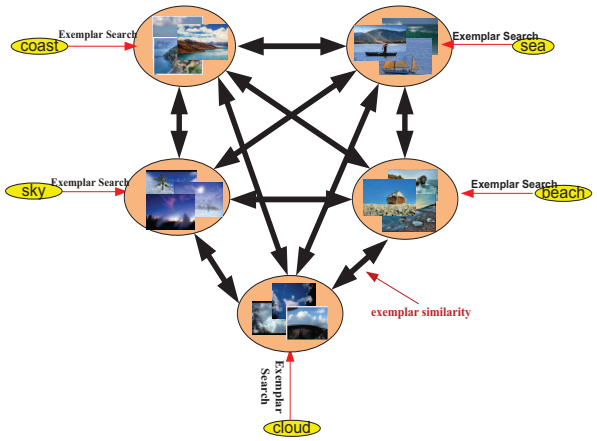


Figure 6: The exemplar similarity between tags is computed based on their representative image collections.

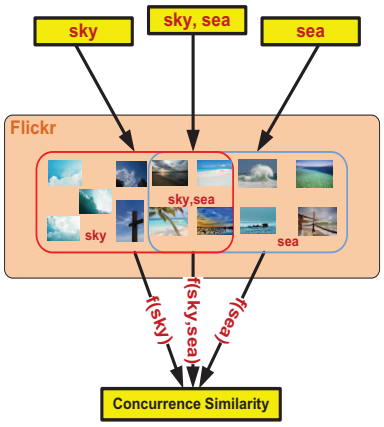


Figure 7: The concurrence similarity between two tags is estimated based on their concurrence information by performing search on Flickr.

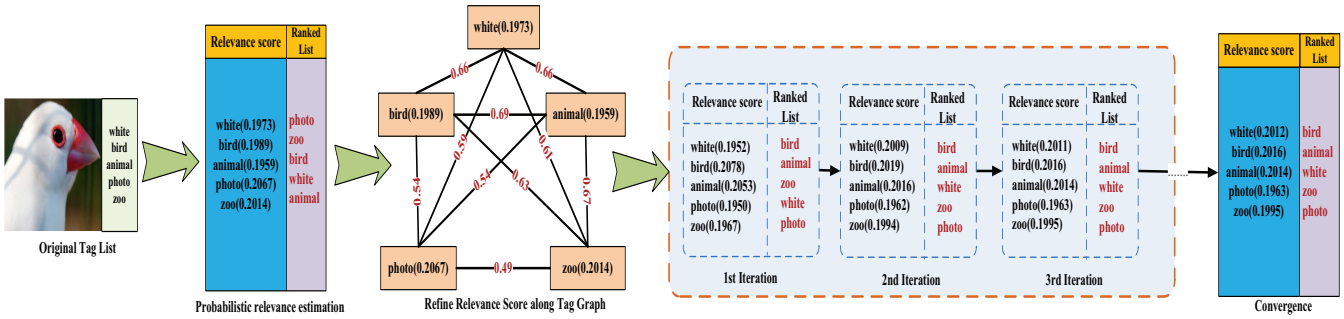


Figure 8: An example of random walk-based relevance refinement process.

PROOF. We re-write Eq. 11 in matrix form

$$\mathbf{r}_k = \alpha \mathbf{P}\mathbf{r}_{k-1} + (1-\alpha)\mathbf{v} \tag{12}$$

and thus we have

$$\mathbf{r}_\pi = \lim_{n\to\infty} (\alpha\mathbf{P})^n \mathbf{r}_0 + (1-\alpha)(\sum_{i=1}^{n}(\alpha\mathbf{P})^{i-1})\mathbf{v}. \tag{13}$$

Note that transition matrix $\mathbf{P}$ has been row normalized to 1. For $0 < \alpha < 1$, there exists $\gamma < 1$,such that $\alpha \leq \gamma$, and we can derive that

$$\begin{aligned}
\sum_j (\alpha\mathbf{P})_{ij}^n &= \sum_j \sum_k (\alpha\mathbf{P})_{ik}^{n-1} (\alpha\mathbf{P})_{kj} \\
&= \sum_k (\alpha\mathbf{P})_{ik}^{n-1} (\alpha \sum_j \mathbf{P}_{kj}) \\
&= \sum_k (\alpha\mathbf{P})_{ik}^{n-1} (\alpha) \\
&\leq \sum_k (\alpha\mathbf{P})_{ik}^{n-1} (\gamma) \\
&\leq \gamma^n
\end{aligned} \tag{14}$$

Thus the row sums of $(\alpha\mathbf{P})^n$ converges to zero. Then according to Eq. 13 we have

$$\mathbf{r}_\pi = (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{v} \tag{15}$$

This is the unique solution. □

Fig. 8 illustrates an example, from which we can see that how the random walk process improves the original probabilistic relevance scores. Despite tag ranking result obtained from the probabilistic relevance estimation is not satisfactory, the iterations of random walk over tag graph are able to produce perfect tag ranking result.

3. PERFORMANCE EVALUATION

3.1 Experimental Settings

All the experiments in this work are conducted on a dataset comprising 50,000 images collected from Flickr. We select ten most popular tags, including *cat*, *automobile*, *mountain*, *water*, *sea*, *bird*, *tree*, *sunset*, *flower* and *sky*, and use them as query keywords to perform tag-based search with ranking by interestingness option. Then the top 5,000 images are collected together with their associated information, including tags, uploading time, etc. for each query. We notice that the collected tags (more than 100,000 unique tags) in this way are rather noisy and many of them are misspelling or meaningless words. Hence, a pre-filtering process is performed for these tags. We match each tag with the entries in a Wikipedia thesaurus, and only the tags that have a coordinate in Wikipedia are kept. In this way, 13,330 unique tags in total are obtained. For each image, we extract 353-dimensional features, including 225-dimenaional block-wise

color moment features generated from 5-by-5 partition of the image and a 128-dimensional wavelet texture features.

We use $NDCG$ [21] as the performance evaluation measure. 10,000 images are randomly selected from our Flickr set for labeling by five persons. For each image, each of its tags is labeled as one of the five levels: Most Relevant (score 5), Relevant (score 4), Partially Relevant (score 3), Weakly Relevant (score 2), and Irrelevant (score 1). Given an image with ranked tag list $t_1, t_2, \ldots, t_n$, the $NDCG$ is computed as

$$N_n = Z_n \sum_{i=1}^{n} (2^{r(i)} - 1)/log(1+i) \quad (16)$$

where $r(i)$ is the relevance level of the ith tag and Z_n is a normalization constant that is chosen so that the optimal ranking's $NDCG$ score is 1. After computing the $NDCG$ measures of each image's tag list, we can average them to obtain an overall performance evaluation of the tag ranking method. The radius parameter σ in Eq. 5 and Eq. 6 is set to the median value of all pair-wise Euclidean distances between images, and the parameter N in exemplar similarity computation in Eq. 6 is empirically set to 50. The parameters λ and α in Eq. 9 and Eq. 12 are set to 0.8 and 0.5 respectively and their sensitivities will be analyzed in detail in section 3.2.

3.2 Experimental Results

We compare our proposed tag ranking strategy with the following two methods in terms of average $NDCG$.

1. Probabilistic Tag Ranking (PTR). In this method, we use the relevance score obtained in Eq. 4 without implementing random walk among tags;

2. Random Walk-based Tag Ranking (RWTR). We conduct random walk along the tag correlation graph with $v_i = 1/n$, i.e., do not utilize probabilistic relevance estimations.

Our proposed approach can be viewed as a combination of the above two methods. The experimental results are shown in Fig. 9. We have also computed the $NDCG$ of the original tag lists as baseline. From the results we can see that all the three methods can bring better order to the tags, and the combination method outperforms PTR and RWTR. Fig. 10 illustrates several exemplary results, from which we can clearly see the tag ranking lists are better than the original ones in terms of relevance ordering.

As shown in Fig. 2, only less than 10% of the images have their most relevant tag at the top position in their attached tag list. We also illustrate the same histogram for the ranked tag list, as shown in Fig. 11. We can clearly see that our tag ranking approach has successfully promoted the most relevant tags. Nearly 40% images have their most relevant tags ranked at top.

We then further conduct experiments to analyze the sensitivity of our tag ranking scheme with respect to the two parameters λ (see Eq. 9) and α (see Eq. 12). The parameter λ is a trade-off to balance the contributions of concurrence-based similarity and exemplar-based similarity in tag graph, and the parameter α controls the impact of probabilistic relevance scores in the random walk process.

First we simply set α to 0.5 and range λ from 0 to 1.0. Fig. 12(a) illustrates the results. From the results we can see

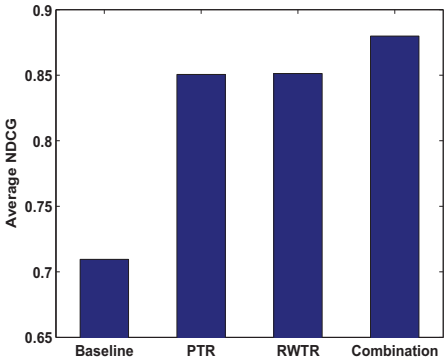


Figure 9: Performance of different tag ranking strategies. Baseline is the average $NDCG$ of the original tag lists and the others are average $NDCG$ values after performing tag ranking with different methods, where PTR is Probabilistic Tag Ranking, PWTR is Random Walk-based Tag Ranking, and the last one is their combination.

that basically setting λ in $(0, 1)$ is better than setting λ to 0 or 1 individually. This demonstrates the complementary nature of the concurrence similarity and exemplar similarity. Interestingly, we find that the optimal performance is achieved when the weighting parameter λ for exemplar similarity is set to 0.8, which confirms the significance of visual modality for our tag ranking. We then set λ to 0.8 and range α from 0 to 1. As can be observed, the performance curve is smooth when α varies in a wide range $[0.3, 0.8]$. We can also see that the result is always better than PTR and RWTR when α ranges from 0.1 to 0.9. This indicates the robustness of our approach. According to the results, the optimal values of λ and α are around 0.8 and 0.5 respectively, and thus in the tag ranking-based applications introduced in the next section we will empirically adopt these settings to generate tag ranking results.

4. APPLICATIONS

In this section, we introduce three potential application scenarios of tag ranking:

1. Tag-based image search, in which the image ranking list is generated based on the given tag's position in these images' ranked tag lists.

2. Tag recommendation. For a given image, we provide the most relevant tags of its neighbors as recommendation.

3. Image group recommendation. Given an image, we use the top tags in the ranked tag list to search for possible groups for sharing purpose.

4.1 Tag-Based Image Search

As introduced in Section 1, currently Flickr provides two tag-based image search services that rank images based on interestingness and uploaded time, respectively. Based on the tag ranking results, we develop a method to rank search results based on their relevance, which will be useful in most

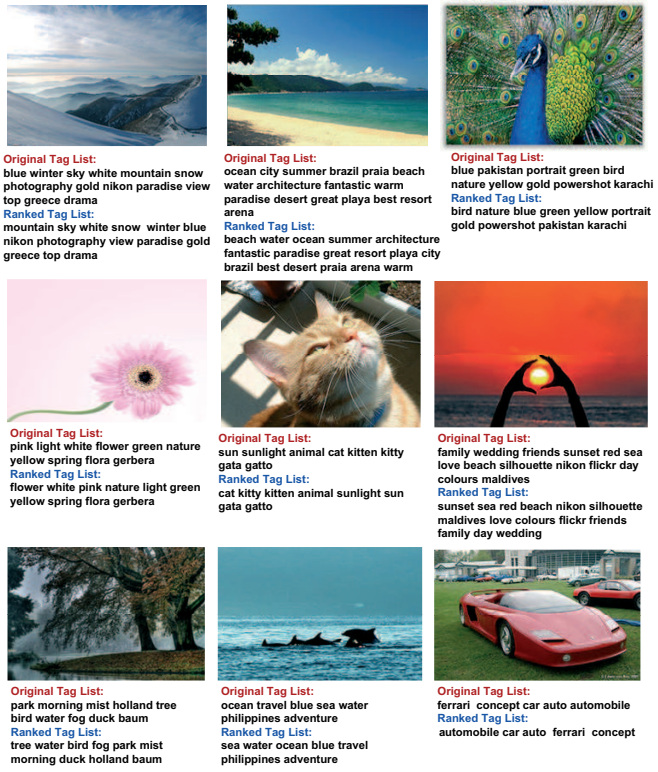


Figure 10: Several exemplary tag ranking results.

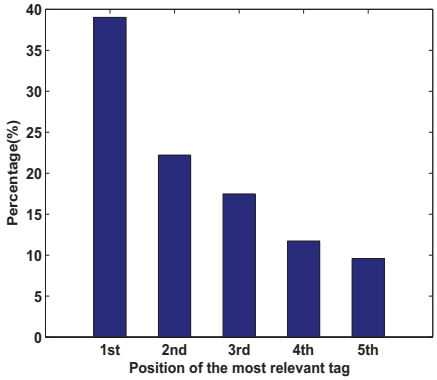
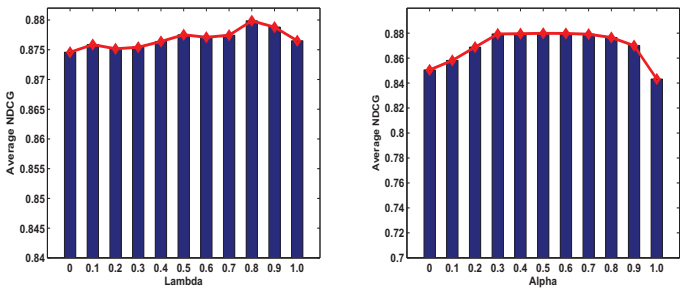


Figure 11: The percentage of images that have their most relevant tag appear at n-th position in the associated tag list after performing tag ranking, where $n = 1, 2, ..., 5$.



(a) (b)

Figure 12: The performance curves of our tag ranking method with respect to the parameters λ and α.

of the situations when users want to search for information about a topic.

Given the query tag, we estimate the relevance levels of images based on the position of the query tag in each image's tag ranking list. Generally, the more advanced position the query tag is located in an image's tag ranking list, the more relevant the image will be with the query. Fig. 13 illustrates an example. The query tag "dog" lies at the first and second positions in the tag lists of (a) and (b) respectively, and we can see that obviously (a) is more relevant to the query than (b).

(a) dog tree (b) tree dog

Figure 13: The tag positions can be used to identify the relevance of images with respect to the query. The query "dog" locates at first and second position in (a) and (b) respectively and we can clearly see that (a) is more relevant than (b) for query "dog".

Our image ranking method is designed as follows. Given a query tag q, denote by τ_i the rank position of q in the ranked tag list of image x_i. Since we only consider the images that contain tag q, we have $\tau_i > 0$. Let n_i denote the number of tags of image x_i, we then estimate relevance score for image x_i as

$$r(x_i) = -\tau_i + 1/n_i \tag{17}$$

The rationality of Eq. 17 lies on the following two facts:

1. If $\tau_i < \tau_j$, we will always have $r(x_i) > r(x_j)$. This indicates that we assign higher relevance score to the image that contains the query tag at more advanced positions in its ranked tag list.

2. If $\tau_i = \tau_j$, the relationship is decided by n_i and n_j. If $n_i < n_j$, we have $r(x_i) > r(x_j)$. It implies that, if the positions of the query tag are the same for two images, then we will prefer to the image that has fewer tags. This is intuitive since fewer tags indicate the image is more simple and it is thus with higher probability to be relevant with the given query tag.

With the ranking scores obtained in Eq. 17, we rank image search results with the scores in descending order. To evaluate the proposed tag-based search strategy, we conduct experiment on the 50,000 Flickr image collection described in section 3.1 and use the 10 tags, i.e., *cat*, *automobile*, *mountain*, *water*, *sea*, *bird*, *tree*, *sunset*, *flower* and *sky*, as query keywords. We use $NDCG$ as the performance evaluation measure. Each image is labeled as one of four levels: Most Relevant (score 4), Relevant (score 3), Weakly Relevant (score 2) and Irrelevant (score 1). We compare our method with the following three methods: (1)

Figure 14: The top results of query "bird" using our image ranking strategy.

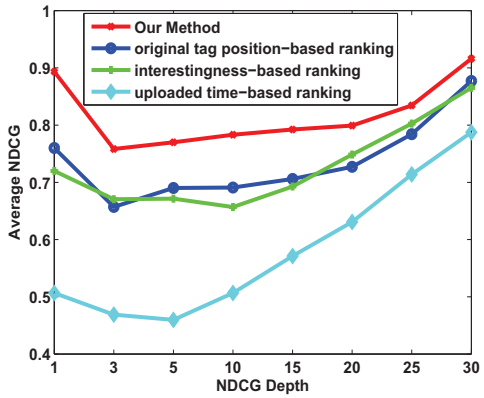


Figure 15: $NDCG$ comparison with varied depths using different ranking methods for tag-based image search.

interestingness-based ranking; (2) uploading time-based ranking; and (3) original tag position-based ranking. The first two methods are both the services provided by Flickr website, which generate image ranking lists based on their interestingness and uploading time records, respectively. In the third method, we use Eq. 17 to generate ranking score with the original tag lists, i.e., do not perform tag ranking. Fig. 14 illustrates the first 12 results of query "bird" using our method. We can compare it with Fig. 3 which illustrates the top results obtained using interestingness-based ranking, and we can clearly see that our top results are more relevant. The average $NDCG$ results with different depths are illustrated in Fig. 15, and the results demonstrate that our method can achieve better search performance, in terms of relevance ranking with respect to the query tag.

4.2 Tag Recommendation

Based on the tag ranking results, we propose a content-based tag recommendation method. As stated in section 5, current tag recommendation approaches can be categorized into automatic and semi-automatic approaches. Our method is one of the automatic approaches and the recommendation process does not require users to provide initial tags. The proposed method is as follows. Given an image,

	Prec@1	Prec@5	Prec@all
Original(Baseline)	0.5858	0.4980	0.4980
Recommendation	0.7255	0.5799	0.5772
Improvement(%)	23.9	16.5	15.9

Table 1: Performance of tag recommendation

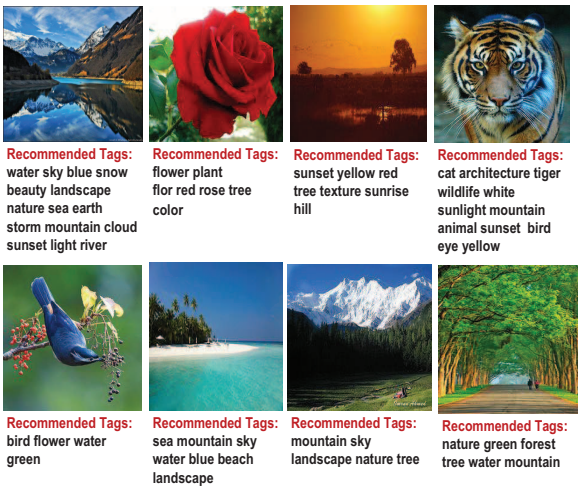


Figure 16: Tag recommendation example.

we first find its K nearest neighbors from the image dataset. Then we collect the top m tags of each neighboring image according to its tag ranking list. As a result, $m \times K$ tags are collected in total. The unique tags in the collection are then recommended to the users, sorted according to the occurrence frequency in the collection.

To demonstrate the effectiveness of the tag recommendation method, we randomly select 500 images from Flickr collection to perform tag recommendation. We set $m = 2$ and $K = 10$ empirically. We also invite people to label the recommended tags as "relevant" or "irrelevant". We use precision as the performance evaluation measure. The average precisions with different depths are illustrated in Table 1. From the table we can see that our tag recommendation results are surprisingly good. The precision is even higher than that of the tags provided by Flickr user (the precisions of the original tags and recommended tags are 0.4980 and 0.5772, respectively). Fig. 16 illustrates some sample images and their recommended tags.

4.3 Group Recommendation

Groups in Flickr are self-organized communities with common interests. Users can add their images to suitable groups such that they can be more easily accessed. The group information is thus useful for users to share and browse images. However, nowadays there are a large number of groups in Flickr and it is not easy for general users to find a suitable group for his photos. Therefore, a group recommendation service is highly desired.

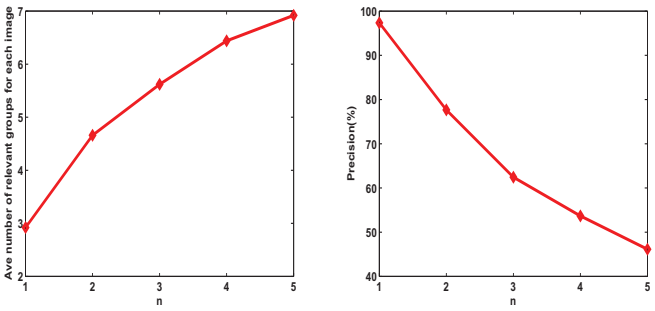
We propose a group recommendation approach based on the tag ranking. Since groups are usually titled and described with words corresponding to the image content in its image pool, we can use the top tags of an image as query keywords to search for its potentially suitable groups. Since

bird nature wildlife black flight action

Tags	Recommended Groups
bird:	Birds and Wildlife UK | Birds Photos | British Birds
nature:	Nature's Beauty | The World of Nature | Arizona Nature
wildlife:	we love wildlife | California Wildlife | The Wildlife Photography

Figure 17: An example for group recommendation. Based on the tag ranking results, we use the first three tags of the given image, i.e., *bird*, *nature* and *wildlife* to search for suitable groups, and we can find a series of possible groups.

the top tags in the ranked tag list are the keywords that can best describe the visual content of the query image, the group will be found with high probability. Fig. 17 shows a typical image and its group recommendation results. The top three tags in the ranked tag list are used as keywords to find the groups which contain these tags. After recommending the potential groups, users can select the preferred groups from the recommended ones.



(a) (b)

Figure 18: Performance of group recommendation with different n. (a) illustrates the average numbers of relevant recommended groups and (b) illustrates the recommendation precisions.

To objectively evaluate the performance of group recommendation, we randomly select 50 images from our Flickr collection. For each image, we use the top n tags in the tag ranking list to search for groups. The top three groups in each search result are collected[4] and thus we obtain $3n$ recommended groups in total. The recommended groups for each image are then manually labeled as "relevant" or "irrelevant" according to whether the image can be categorized into the groups for evaluation purpose. We calculate group recommendation precision for each image and average them as final evaluation measure. n varies from 1 to 5, and Fig. 18 illustrates the average results. Fig. 18(a) shows the average number of relevant groups for an image and (b) illustrates the recommendation precision, i.e., the percentage of relevant ones in the recommended groups. From the figure we can see that the number of relevant groups keeps increasing when n increases, but the recommendation precision actually decreases. This can be easily understood, since using more tags can always find more groups and there will be more relevant ones, but the group recommendation precision will be worse than only using one or two most relevant tags. If we only use the top tag for search, we can see that the recommended groups will nearly be all relevant, and this demonstrates the effectiveness of our approach. In practical application, we can establish n by compromising the number of recommended groups and the recommendation precision or providing several options for user customization.

[4]Currently Flickr supports four text-based group search strategies, including "ordering by most relevant", "ordering by most recent activity", "ordering by group size" and "ordering by date created". We adopted the first strategy.

5. RELATED WORK

The tags that describe the content of images can help users easily manage and access large-scale image repositories. With these metadata, the manipulations of image data can be easier to be accomplished, such as browsing, indexing and retrieval. Extensive research efforts have been dedicated to automatically annotating images with a set of tags [6, 7, 8, 9]. These methods usually require a labeled training set and then learn models with these data based on low-level features, and then new unlabeled images can thus be annotated using these models. Although a lot of encouraging results have been reported, the performance of these approaches are still far from satisfactory for practical large-scale application due to the semantic gap between tags and low-level features. Manual tagging that allows users to provide image tags by themselves is an alternative approach. It of courses adds users' labor costs in comparison with automatic tagging, but it will provide more accurate tags. Manual tagging has widely been adopted in image and video sharing websites such as Flickr and Youtube, and the popularity of these websites has demonstrated its rationality. However, as introduced in Section 1, the user provided tags are orderless and this significantly limits their applications. Asking users to manually order the tags is obviously infeasible since it will add much more labor costs for users. This difficulty motivates our work. To the best of our knowledge, this is the first work that aims to bring orders to image tags.

Flickr, as the most popular social image sharing web site, has been intensively studied in recent years, especially on its tagging characteristic. Ames et al. [2] have explored the motivation of tagging in Flickr website and they claim that most users tag images to make them better accessible to the general public. Sigurbjörnsson et al. [10] have provided the insights on how users tag their photos and what type of tags they are providing. They conclude that users always tag their photos with more than one tag and these tags span a broad spectrum of the semantic space. Kennedy et al. [3] have evaluated the performance of the classifiers trained with Flickr images and associated tags and demonstrate that tags provided by Flickr users actually contain many noises. Li et al. [11] have proposed to learn tag relevance to boost tag-based social image retrieval. Yan et al. [12] have proposed a model that can predict the time cost of manual image tagging. Different tag recommendation methods have been proposed that aim to help users tag

images more efficiently [10, 13, 14]. The existing tag recommendation methods can be categorized into automatic and semi-automatic approaches. Semi-automatic tag recommendation requires users to provide one or several tags firstly and then conduct the recommendation accordingly. The methods proposed by Sigurbjörnsson et al. [10] and Wu et al. [13] belong to this category. Automatic recommendation is usually accomplished by exploiting the image content. Chen et al. [14] proposed an automatic tag recommendation approach that directly predicts the possible tags with models learned from training data. This method thus can only recommend the tags from a predefined set. Our proposed tag recommendation method belongs to the automatic category, but it does not need models of tags and can avoid the difficulty in [14]. Chen et al. [14] also proposed to use the predicted tags to search for groups as recommendation groups for the given image, but this method heavily relies on the performance of tag prediction. In our proposed group recommendation method, we use the most relevant tags from those provided by users for group search and the performance can thus be better guaranteed.

6. CONCLUSIONS

In this paper, we have shown that the tags associated with Flickr images are almost orderless, which limits the effectiveness of the tags in many related applications. We thus propose an approach to rank the tags for each image according to their relevance levels. Our experimental results demonstrate that the proposed method can order the tags according to their relevance levels. We then propose three application scenarios which can benefit from tag ranking, including tag-based search, tag recommendation and group recommendation. Encouraging results are reported, which demonstrates the effectiveness of the proposed tag ranking approach.

It is worth noting that although we have only used Flickr data in this work, the proposed tag ranking method is a general approach and can be applied for other data sources (e.g., Youtube for video tag ranking). We believe this work can provide new facilities and opportunities for social media tagging services.

7. REFERENCES

[1] Flickr. *http://www.flickr.com.*

[2] M. Ames and M. Naaman. Why We Tag: Motivations for Annotation in Mobile and Online Media. In *Proceeding of the SIGCHI Conference on Human Factors in Computing System*, 2007.

[3] L. S. Kennedy, S. F. Chang, and I. V. Kozintsev. To Search or To Label? Precdicting the Performance of Search-Based Automatic Image Classifiers. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, 2006.

[4] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based Multimedia Information Retrieval: State of the Art and Challenges. In *ACM Transactions on Multimedia Computing, Communications and Applications*, 2006.

[5] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain. Content based Image Retrieval At the End of the Early Years. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[6] J. Li and J. Z. Wang. Real-Time Computerized Annotation of Pictures. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[7] F. Monay and D. G. Perez. On Image Auto-annotation with Latent Space Modeling. In *Proceeding of 10th ACM International Conference on Multimedia*, 2003.

[8] G. Sychay, E. Y. Chang and K. Goh. Effective Image Annotation via Active Learning. In *IEEE International Conference on Multimedia and Expo*, 2002.

[9] R. Shi, C. H. Lee and T. S. Chua. Enhancing Image Annotation by Integrating Concept Ontology and Text-based Bayesian Learning Model. In *Proceeding of 14th ACM International Conference on Multimedia*, 2007.

[10] B. Sigurbjörnsson and R. V. Zwol. Flickr Tag Recommendation based on Collective Knowledge. In *Proceeding of ACM International World Wide Web Conference*, 2008.

[11] X. R. Li, C. G. M. Snoek and M. Worring. Learning Tag Relevance by Neighbor Voting for Social Image Retrieval. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, 2008.

[12] R. Yan, A. Natsev and M. Campbell. A Learning-based Hybrid Tagging and Browsing Approach for Efficient Manual Image Annotation. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[13] L. Wu, L. J. Yang, N. H. Yu and X. S. Hua. Learning to Tag. In *Proceeding of ACM International World Wide Web Conference*, 2009.

[14] H. M. Chen, M. H. Chang, P. C. Chang, M. C. Tien, W. H. Hsu and J. L. Wu. SheepDog-Group and Tag Recommendation for Flickr Photos by Automatic Search-based Learning. In *Proceeding of 15th ACM International Conference on Multimedia*, 2008.

[15] E. Parzen. On the Estimation of a Probability Density Function and the Mode. In *Annals of Mathematical Statistics*, 1962.

[16] B. Liu. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. *Springer, December*, 2006.

[17] R. Cilibrasi and P. M. B. Vitanyi. The Google Similarity Distance. In *IEEE Transactions on Knowledge and Data Engineering*, 2007.

[18] W. H. Hsu, L. S. Kennedy and S. F. Chang. Video Search Reranking through Random Walk over Document-Level Context Graph. In *Proceeding of 14th ACM International Conference on Multimedia*, 2007.

[19] Y. S. Jing and S. Baluja. VisualRank: Applying PageRank to Large-Scale Image Search. In *Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[20] J. Y. Pan, H. J. Yang, C. Faloutsos and P. Duygulu. Gcap: Graph-based Automatic Image Captioning. In *International Workshop on Multimedia and Document Engineering*, 2004.

[21] K. Jarvelin and J. Kekalainen. Cumulated Gain-Based Evaluation of IR Techniques. In *ACM Transcations on Information System*, 2002.