Which Tags Are Related to Visual Content?

Yinghai Zhao[1], Zheng-Jun Zha[2], Shanshan Li[1], and Xiuqing Wu[1]

[1] University of Science and Technology of China, Hefei, Anhui, 230027, China
[2] National University of Singapore, Singapore, 639798
{yinghai,ssnl}@mail.ustc.edu.cn,
junzzustc@gmail.com, xqwu@ustc.edu.cn

Abstract. Photo sharing services allow user to share one's photos on the Web, as well as to annotate the photos with tags. Such web sites currently cumulate large volume of images and abundant tags. These resources have brought forth a lot of new research topics. In this paper, we propose to automatically identify which tags are related to the content of images, i.e. which tags are content-related. A data-driven method is developed to investigate the relatedness between a tag and the image visual content. We conduct extensive experiments over a dataset of 149,915 Flickr images. The experimental results demonstrate the effectiveness of our method.

Keywords: Flickr, tag, content-relatedness, visual content.

1 Introduction

Online photo-sharing services, such as Flickr [1] and Photobucket [2], encourage internet users to share their personal photos on the web, as well as to annotate the photos with tags (*i.e.*, keywords). Take Flickr for example, it currently cumulates around four billion images as well as billions of tags [1].

These gigantic volume of social tagged images have brought forth many novel research topics. For example, Sigurbjörnsson et al. [7] proposed a tag recommendation strategy based on tag concurrence analysis, while Wu et al. [9] proposed to recommend tags by taking both tag concurrence and image visual content into account. The social tagged images are also used to aid image search. Liu [10] and Li [4] proposed to compute the relevance between tags and images, which in turn facilitated the image search. Although these works investigated the usage of the social images and tags. The facets of tags have been seldom studied.

As reported in [3], the tags associated with images are mainly to describe the image contents and provide other information, such as time stamp, location, and subjective emotion. We argue that the automatic identification of tags which are content-related can aid more intelligent use of the social images and tags and thus facilitate the researches and applications over these resources. To do this, we propose a data-driven method to analyze the relatedness between the tags and the content of the images.

Specifically, different tags have different prior probabilities to be content-related because of the semantic nature of tags. For example, tag *"flower"* is more

S. Boll et al. (Eds.): MMM 2010, LNCS 5916, pp. 669–675, 2010.

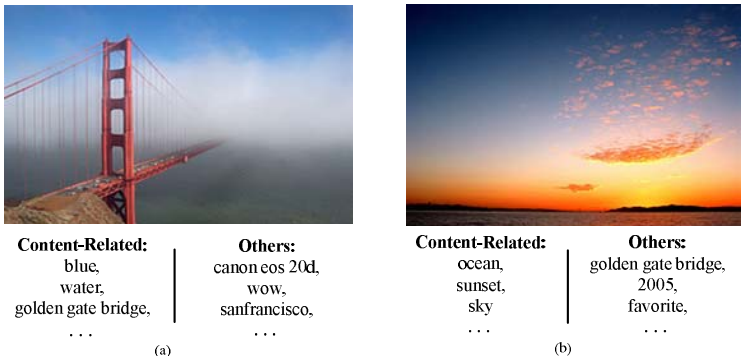

Fig. 1. Two Flickr images and the tags labeled to them

probable to be content-related than *"music"*. Moreover, the content-relatedness between the same tag and the associated images may vary with different images. Fig. 1 shows two images which are annotated with tag *"golden gate bridge"*. However, Fig. 1(a) is a photo of the bridge while (b) is a sunset picture maybe taken on the bridge. Thus, the tag "*golden gate bridge*" is content-related to Fig. 1(a), but is not content-related to Fig. 1(b). Therefore, we investigate the probability that a tag is related to the content of a specific image from the above two aspects.

The rest of this paper is organized as follows. Our method is elaborated in Section 2. Then, the evaluation results over Flickr images are reported in Section 3. Finally, we give the conclusion remarks in Section 4.

2 Tag Content-Relatedness Analysis

Intuitively, (1) if one tag is often used by different users to annotate similar images, this tag is widely accepted to be a proper description for some objective aspects of images content, i.e., this tag is content-related in nature. Moreover, (2) if one tag is labeled to an image by one user, and this image is similar to many other images labeled with this tag by different users, this tag is very likely to be content-related to the specific image. Motivated by these two observations, we propose the following method.

Given a tag t and a set of images $\mathcal{X} = \{x_i\}_{i=1}^N$ that are annotated with t, our target is to derive the scores $\mathcal{P} = \{p_i\}_{i=1}^N$ which measure the content-relatedness between tag t and each image x_i. The probability p_i can be represented as $p_i(r = 1|t, x_i, \mathcal{X}\backslash x_i)$, where $r \in \{1, 0\}$ is an indicator of being content-related or not and $\mathcal{X}\backslash x$ denotes all the other images in $\mathcal{X}$ except x. To reduce the tagging bias of single user in $\mathcal{X}$, each user is limited to contribute only one image to $\mathcal{X}$. Thus, x and $\mathcal{X}\backslash x$ are conditionally independent given t. According to Bayes' formula, the probability $p(r = 1|t, x, \mathcal{X}\backslash x)$ can be derived as follows:

$$p(r = 1|t, x, \mathcal{X}\backslash x) = \frac{p(r = 1|t, \mathcal{X}\backslash x)p(r = 1|t, x)}{p(r = 1|t)}, \tag{1}$$

where $p(r = 1|t)$ indicates how likely t is content-related without knowing any prior knowledge, and is set to be uniform over all tags.

Then, Eq.1 can be formulated as:

$$p(r = 1|t, x, \mathcal{X}\backslash x) \propto p(r = 1|t, \mathcal{X}\backslash x)p(r = 1|t, x), \tag{2}$$

where the item $p(r = 1|t, \mathcal{X}\backslash x)$ indicates the prior probability that t is content-related given the social tagged image resources $\mathcal{X}\backslash x$, and $p(r = 1|t, x)$ expresses the likelihood that t is content-related to x. Overall, $p(r = 1|t, x, \mathcal{X}\backslash x)$ gives out the posterior probability of t being content-related to x with the assistance of social tagged images $\mathcal{X}\backslash x$. For simplicity, we denote $p(r = 1|t, \mathcal{X}\backslash x)$, $p(r = 1|t, x)$ and $p(r = 1|t, x, \mathcal{X}\backslash x)$ as $PrCR$, $LiCR$ and $PoCR$, respectively.

2.1 Probability Estimation

Because one object could be presented from different points of view and one image may only show some local parts of the object, we estimate $p(r = 1|t, \mathcal{X}\backslash x)$ through the local visual consistency over all $x_i \in \mathcal{X}\backslash x$. Here, local visual consistency is a measurement of visual similarities between an image and its K-nearest neighbors. The $p(r = 1|t, \mathcal{X}\backslash x)$ can be estimated as:

$$p(r = 1|t, \mathcal{X}\backslash x) = \frac{1}{KN} \sum_{i=1}^{N} \sum_{j=1}^{K} s(x_i, x_j), \tag{3}$$

where $x_j \in \mathcal{X}\backslash x$ is one of the K-nearest neighbors of x_i, and $N{=}|\mathcal{X}\backslash x|$ is the number of images in $\mathcal{X}\backslash x$. Moreover, $s(x_i, x_j)$ denotes the visual similarity between x_i and x_j .

Similarly, the likelihood $p(r = 1|t, x)$ can be evaluated through the local visual consistency of image x with respect to its K'-nearest neighbors in $\mathcal{X}\backslash x$:

$$p(r = 1|t, x) = \frac{1}{K'} \sum_{i=1}^{K'} s(x, x_i), \tag{4}$$

where $x_i \in \mathcal{X}\backslash x$, $i = 1, \cdots, K'$, are the K' nearest neighbors in $\mathcal{X}\backslash x$. For simplicity, we let $K = K'$ in our method.

2.2 Visual Similarity

In this subsection, we show the different definitions of visual similarity measurement $s(.)$. It's widely accepted that global features, such as color moments and GIST [6], are good at characterizing scene-oriented (*e.g.*, "*sunset*"), color-oriented (*e.g.*, "*red*") images, while local features, such as SIFT [5], perform better for object-oriented (*e.g.*, "*car*") images. The fusion of multiple features can achieve better representation for image content. Here, we use three similarity definitions that are based on global, local features, and both global and local features, respectively.

f is a feature vector that may be the concatenation of several kinds of global features extracted from image x, the global visual similarity between two images x_i and x_j can be calculated through Gaussian kernel function as:

$$s_g(x_i, x_j) = \exp(-\frac{||f_i - f_j||^2}{\sigma^2}), \tag{5}$$

where σ is the radius parameter of Gaussian kernel.

To computer local visual similarity, bag-of-visual-words method is adopted here [8]. Each image is represented as a normalized visual word frequency vector of dimension D. Then, the local visual similarity between two images could be calculated through the cosine similarity:

$$s_l(x_i, x_j) = \frac{v_i^T v_j}{||v_i||||v_j||}, \tag{6}$$

where v_i and v_j are visual word representations of x_i and x_j, respectively.

Furthermore, a fused visual similarity is obtained through the line combination of global and local visual similarities:

$$s_c(x_i, x_j) = \alpha s_g(x_i, x_j) + (1 - \alpha) s_l(x_i, x_j), \tag{7}$$

where $0 < \alpha < 1$ is the combination coefficient.

3 Experiments

3.1 Data and Methodologies

For experimental data collection, the 60 most popular Flickr tags in April 2009 are selected as seed queries. For each query, the first 2000 images are collected through Flickr image searching. During this process, only the first image is kept for one user. Tag combination and tag filtering operations are conducted to get the tags with more than one word and remove noises. Afterwards, another 207 most frequent tags are selected as queries for further image collection. Finally, 149,915 images are obtained in all.

The following visual features are used to characterize the content of images:

- global features: 6-dimensional color moment in LAB color space and 100-dimentional GIST feature [6] processed by PCA.
- local features: 128-dimensional SIFT descriptors [5].

For parameter setting, the nearest neighbor number, K and K', are both set to 100, and the size of $\mathcal{X}$ is set to 600. Moreover, the size of the visual word codebook D is set to 5000, and the coefficient α in Eq.7 is set to 0.5.

3.2 How Probable the Tag Is Content-Related

In this experiment, we investigate the prior content-relatedness ($PrCR$) probabilities $p(r = 1|t, \mathcal{X}\backslash x)$ for different tags. The $PrCR$ results calculated with

Table 1. The first 10 and the last 10 tags in the *PrCR* based tag sorting results. The tags in bold are examples being ranked at inappropriate places.

Position	*PrCR_g*	*PrCR_l*	*PrCR_f*	Position	*PrCR_g*	*PrCR_l*	*PrCR_f*
1	**winter**	flowers	flowers	51	italy	australia	trip
2	snow	cat	**winter**	52	wedding	trip	canon
3	blue	dog	snow	53	music	nature	japan
4	flower	food	flower	54	europe	vacation	europe
5	beach	**christmas**	green	55	france	wedding	france
6	green	people	cat	56	nyc	birthday	music
7	flowers	snow	beach	57	paris	canon	spain
8	water	green	food	58	spain	new	wedding
9	sky	flower	blue	59	taiwan	california	california
10	cat	city	water	60	california	taiwan	taiwan

global visual similarity, local visual similarity and fused visual similarity are denoted with $PrCR\_g$, $PrCR\_l$, and $PrCR\_f$, respectively. We sort the 60 Flickr tags according to their $PrCR$ probabilities in descending order. Due to the limited space, only the first 10 and the last 10 tags are listed in Tab.1.

From the results in Tab.1, we observe that:

- Tags whose semantics are related to concrete objects, colors, or scenes are generally ranked at the top of the results, while tags whose semantics are related to locations, abstract concepts or time, are ranked at the bottom.
- The $PrCR\_g$ metric succeeds to promote scenes and colors oriented tags, such as "*beach*", and '*blue*", to the top of the list while the $PrCR\_l$ metric prefers concrete objects related tags, such as "*flowers*", "*dog*".
- $PrCR\_f$ is benefited from the fusion of global and local similarity measurements, and gets the most reasonable result.

3.3 How Probable the Tag Is Content-Related to Specific Image?

In this section, we evaluate the performance of the proposed method in sorting the image tags according to their content-relatedness probability. For evaluation, 400 images are randomly selected with their tags being manually labeled into two levels: "content-related" or not. We measures the tag sorting performance with mean average precision (MAP). MAP is obtained by averaging the AP scores of the sorted tag lists over all the test images.

We compare the performances of five methods: (1) *Baseline* which are the tagging results in the order of user inputting; (2) $LiCR\_g$ which sorts tags based on $p(r = 1|t, x)$ with global visual similarity; (3) $LiCR\_l$ which sorts tags according to $p(r = 1|t, x)$ with local visual similarity; (4) $LiCR\_f$ which sorts tags according to $p(r = 1|t, x)$ with fused similarity; and (5) $PoCP$ which sorts tags according to posterior probability $p(r = 1|t, x, \mathcal{X}\backslash x)$ with fused visual similarity.

The experimental results are shown in Fig. 2 (a). From the results, we can observe that all the last four methods could consistently boost the tag sorting performance comparing to the original input order. Our proposed method $PoCR$

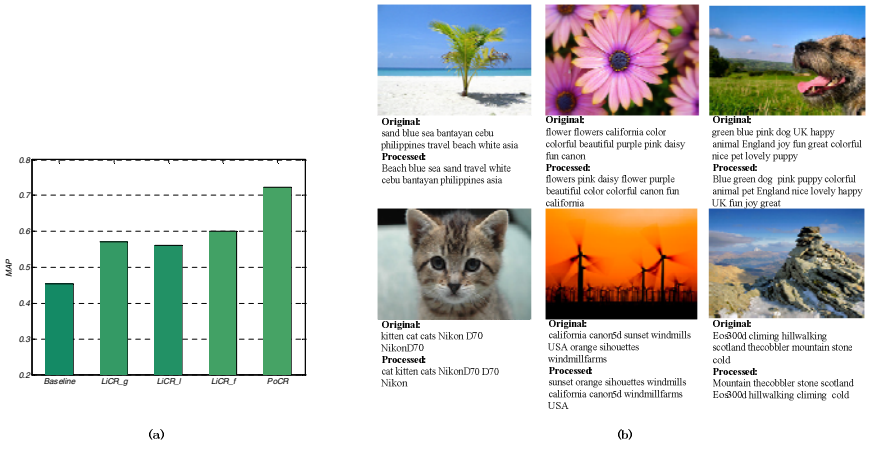

Fig. 2. (a) MAP results of tag content-relatedness analysis for 400 images. (b) Some example results of tag content-relatedness analysis. According to our method, the tags which are content-related to the image are ranked at the top of the processed tag list.

achieves the best performance, and obtains 59.8%, 26.6%, 29.3%, and 20.4% relative improvements compared to *Baseline*, $LiCR\_g$, $LiCR\_l$, and $LiCR\_f$ respectively. We illustrate some example images and their original tag lists, sorted tag lists, in Fig. 2 (b).

4 Conclusion

In this paper, we have firstly shown the fact that tags of Flickr images are not always content-related to images. Then, we propose an data-driven approach to evaluate the probability of a tag to be content-related with respect to an image. It's worth noting that our method requires no model training process and could be scalable to large-scale datasets easily. Experiments on 149,915 Flickr images demonstrate the effectiveness of the proposed method.

References

1. Flickr, `http://www.flickr.com/`
2. Photobucket, `http://photobucket.com/`
3. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI conference on Human factors in computing systems, San Jose, California, USA, pp. 971–980 (2007)
4. Li, X., Snoek, C.G., Worring, M.: Learning tag relevance by neighbor voting for social image retrieval. In: Proceeding of the 1st ACM international conference on Multimedia information retrieval, Vancouver, Canada, pp. 180–187 (2008)
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)

6. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision 42(3), 145–175 (2001)
7. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceeding of the 17th international conference on World Wide Web, Beijing, China, pp. 327–336 (2008)
8. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 1470–1477 (2003)
9. Wu, L., Yang, L., Yu, N., Hua, X.-S.: Learning to tag. In: Proceedings of the 18th international conference on World wide web, Madrid, Spain, pp. 361–370 (2009)
10. Liu, D., Hua, X.-S., Yang, L., Wang, M., Zhang, H.-J.: Tag ranking. In: Proceedings of the 18th international conference on World wide web, Madrid, Spain, pp. 351–360 (2009)