Killing Two Birds With One Stone: Concurrent Ranking of Tags and Comments of Social Images

Boon-Siew Seah  Aixin Sun  Sourav S Bhowmick
School of Computer Science and Engineering,
Nanyang Technological University, Singapore
axsun|assourav@ntu.edu.sg

ABSTRACT
User-generated comments and tags can reveal important visual concepts associated with an image in Flickr. However, due to the inherent noisiness of the metadata, not all user tags are necessarily descriptive of the image. Likewise, comments may contain spam or chatter that are irrelevant to the image. Hence, identifying and ranking relevant tags and comments can boost applications such as tag-based image search, tag recommendation, etc. In this paper, we present a lightweight visual signature-based model to concurrently generate ranked lists of comments and tags of a social image based on their joint relevance to the visual features, user comments, and user tags. The proposed model is based on sparse reconstruction of the visual content of an image using its tags and comments. Through empirical study on Flickr dataset, we demonstrate the effectiveness and superiority of the proposed technique against state-of-the-art tag ranking and refinement techniques.

ACM Reference Format:

1 INTRODUCTION
In social image platforms like Flickr and Instagram, users may annotate an image with tags as well as add comments related to multiple aspects of an image. In particular, more than 90% of images in NUS-WIDE dataset has received at least one comment [2]. [1] A subset of these comments may serve as a potential source of important information about the image. However, these comments are often riddled with noise and irrelevant chatter, making it hard for any automated technique to correlate them with the visual content or context of an image.

Consider Figure 1 depicting an image from Flickr along with original comments and tags in Figures 1(a) and 1(b), respectively. We can make the following key observations. (a) Only a subset of the
tags (e.g., fisheye, breakdancing) is interesting to a typical user as highlighted by the comments. Tags such as d200 and 2006 are not brought up in discussions, indicating that they are of little interest to viewers. Furthermore, some of the comments are not relevant to the visual content of the image (e.g., “From the samples you can see the scientific answer is pretty darned wide”). (b) Based on the discussion in comments, the visual concepts that capture users’ attention include the visual effect (e.g., fisheye) demonstrated in the photo, the scene captured by the photo (e.g., breakdancing), and the emotional effect arise from viewing the image (e.g., funny).

Given such disparate collections of tags and comments, how can we identify and rank them according to their relevance to the visual content of the image? We believe that the answer to this question benefits several applications in social image search, particularly in building superior image ranking model and search result snippet generation. Consequently, in this paper, given an image we leverage on its visual features, user comments, and tags to concurrently rank tags and comments according to their relevance to the visual content of the image. Specifically, we aim to simultaneously answer the following two questions: (a) Which visual concepts (represented by tags) in an image capture most users’ attention and discussion?
We denote a social image as a tuple \( \langle \text{work}, \text{viewers} \rangle \) where the work is represented by a set of visual features \( \{v_1, v_2, \ldots, v_m\} \) and the viewers are relevant to the content of an image. We present a novel visual signature-based model for jointly ranking tags and comments. The model not only incorporates the semantic and visual properties of the tags associated with a social image, but also evaluates the user comments to generate superior quality results. A distinguishing feature of our model is that it is lightweight in nature. Specifically, it produces superior ranking without leveraging on deep learning (and expensive training process). By applying our proposed technique to real-world Flickr images, we show its effectiveness and significant improvement of performance over existing methods that rely only on visual and semantic properties of tags and images.

2 RELATED WORK

There have been several efforts related to tag relevance learning (i.e., determining the effectiveness of a tag in describing the visual content of the tagged image) and using it to rank or refine tags. Li et al. [6] proposed to learn tag relevance by visual nearest neighbor voting. The authors in [7] used neighbor-voting as the first step and then applied random-walk to further refine the learned tag relevance. Wu et al. formulate the problem of co-ranking tags and images into a Bregman divergence optimization framework [13]. Feng et al. [3] improved tag ranking by learning from limited training image dataset. The relevance learning is also related to the tag refinement task where less-relevant user-assigned tags may be removed while more-relevant tags to the image content are suggested [4]. Recently, [10] used deep learning-based image classification and object detection techniques to improve tag recommendation. However, none of these efforts focuses on ranking comments and tags concurrently.

There is also increasing attention on using user comments to assist in social image retrieval. Wang et al. [12] utilizes comments together with other textual features for sentiment analysis of social images. Comments are also used to predict what viewer affect concepts (e.g., "delicious" and "hungry") will be evoked after viewing an image with affect tags (e.g., "yummy food") [1]. Momeni et al. proposed an approach that can rate the quality of a Flickr comment [8]. More recently, they proposed to rank comments by enriching them with multiple semantic facets [9]. However, without the assistance of tags, it is extremely difficult to determine which comments are relevant to the content of an image.

3 TAG AND COMMENT RANKING PROBLEM

We denote a social image as a tuple \((v, t, c)\). The visual content of an image is represented by a set of visual features \(v\). Users may add comments to an image. By abusing the notation of lists, we represent these comments by the list \(c = [c_1, c_2, \ldots, c_m]\) where the comments are ordered by their time of posting. Each comment \(c_j \in c\) is modeled as a bag-of-words: \(c_j = \{w_1, w_2, \ldots, w_n\}\). We denote the frequency of a word \(w\) appearing in a comment \(c_j\) as \(\text{count}(w, c_j)\). The set of tags associated with an image is denoted as \(t = \{t_1, t_2, \ldots, t_n\}\). We assume that the tags are high-level semantic concepts assigned by image uploaders or annotators. Hence, in this paper we use tag and concept interchangeably. Furthermore, by word matching, we assume that each comment \(c_j\) is associated with a concept set, denoted by \(\text{concept}(c_j) = \{t_1, t_2, \ldots\}\). For example the comment “this is a cute cat” is associated with the concept set (“cute”, “cat”) if both cute and cat have been used as tags. If a comment does not match any concept (e.g., wow!), then its associated concepts is an empty set \(\text{concept}(c_j) = \emptyset\). Given a social image \((v, t, c)\), the goal of our research is to by rank tags and comments them according to their relevance to the image visual content.

4 VISUAL SIGNATURE-BASED MODEL

Intuitively, the sets of tags and comments of an image describe a set of visual signatures of the image. In this section, we introduce a novel visual signature-based strategy: (i) to select a sparse subset of comments sufficient to reconstruct the visual signature of the tags, and (ii) to select a sparse subset of tags sufficient to reconstruct the visual signature of the comments.

Visual Signatures of Words and Tags. We first introduce the notion that a word or tag may carry visual information. Consider, for example, the line tag. By analyzing all images annotated with the line tag, one may find that it is significantly associated with the edge direction visual features. Such word will be useful in representing images having strong edge directionality features. We refer to such word (tag) as visually active. Visually active words form the building blocks toward the reconstruction of an image.

Formally, the visual signature of a visually active word is represented as a vector. Given an image \((v, t, c)\) and a word \(w\), let \(v^w\) be a vector of weights for the visual feature vector \(v\). This vector is a representation of significant visual features that are associated with this word. To evaluate the significance of a visual feature \(v^x\), we use the following ratio: \(\chi^2 = \frac{(v^x - E[v^x])^2}{E[v^x]}\). Then, we consider a visual feature significant when it’s ratio exceeds a user-defined threshold \(\theta\). If \(v^c_i\) is a significant visual feature of \(w\), then \(v^c_i > \theta\); otherwise \(v^c_i = 0\).

Visual Information Representation. We now extend the idea to represent an image \((v, t, c)\) via a subset of its comments \(c\) and tags \(t\). Given tags \(t\), the visual information of the image supported by the tags is defined as: \(y_t = v \odot \frac{1}{Z} \sum_{x \in t} v^x\) where \(\odot\) is the entrywise product operation. Here \(y_t\) represents the visual information of the image that can be represented by the tags \(t\). Likewise, the visual information described by the entire corpus of comments is defined as \(y_c = v \odot \frac{1}{Z} \sum_{c \in c} \sum_{x \in c} v^x \cdot \text{count}(x)\) where \(Z = \sum \text{count}(x)\) normalizes the vector. Here \(\text{count}(x)\) is the frequency of the concept occurring in the current image’s comments and tags.

In the visual signature-based model, we aim to identify the followings: (a) A subset of comments \(c_1 \subset c\) such that \(y_{c_1}\) is sufficiently similar to the tags visual representation vector \(y_t\). (b) A subset of
tags $t_1 \subset t$ such that $y_{t_1}$ is sufficiently similar to comments visual representation vector $y_c$.

The joint reconstruction identifies a subset of tags and a subset of comments that are relevant to each other with respect to the image visual features. The set of comments captures a subset of significant visual features of the image. Similarly, the set of tags captures different visual signatures of the image. Then the goal of reconstruction is to select a sufficient subset of comments and tags such that (a) they capture most of the visual signatures of the image, and (b) the visual signatures captured using comments and tags are “similar” to each other.

To achieve this goal, we assign weights to each comment $c$. The weight vector for selecting representative comments is represented by $w_c$, and only positively weighted comments are selected. At the same time, the weight vector $w_t$ selects representative tags. Then, we find the appropriate weights $w_c$ and $w_t$ by solving the following optimization problem:

$$
\arg\min_{w_c, w_t} \| y_t - X_c w_c \|^2_2 \quad \text{arg}\min_{w_c, w_t} \| y_c - X_t w_t \|^2_2
$$

The goal is to minimize the Frobenius-norm reconstruction errors of both tags and comments visual information vectors. The selected comments can reconstruct closely the visual signatures of the selected tags, and vice versa.

We introduce regularization that penalizes weight differences between similar tags and words. This is facilitated by using a graph structure based on the generalized Lasso problem [11]. To penalize weight differences between similar tags, we construct a tag-tag constraint graph ($G_t, E_t$) where $V_t$ are the tag nodes and we add an edge $(i, j) \in E_t$ if $\text{sim}(i, j)$ is greater than a cut-off threshold $\delta$. Given the graph, the weight difference penalty function is given by:

$$
L(G_t) = \begin{cases} 
|w_t(i) - w_t(j)| & \text{if } (i, j) \in E_t \\
0, & \text{otherwise}
\end{cases}
$$

Similarly, we can construct a comment-comment constraint graph $L(G_c)$ using the above approach. The optimization problem then is defined as follows:

$$
\arg\min_{w_c, w_t} \| y_c - X_t w_t \|^2_2 + \lambda L(G_t) + \beta \|w_t\|_1
$$

where $\lambda$ specifies the penalty effect of the constraint graphs and $\beta$ specifies the sparsity penalty. With the weight vectors $w_c, w_t$, we obtain the set of interesting comments and tags by choosing $x$ whenever $w(x) > 0$. The above equations can be solved using the path solution for the generalized Lasso problem [11]. Then, the comments and tags can be ranked by the generalized weight values $w(x)$.

5 EXPERIMENTS

We evaluate the proposed model on NUS-WIDE corpus containing more than 269K Flickr images [2]. We crawled their tags and comments through Flickr API. Each image is represented by the followings: (a) a visual feature vector describing the visual content of the image, (b) user comments, and (c) user tags. The visual features are provided by the dataset. In this study, we use all tags associated with an image without filtering them. Note that the size of the image collection does not impact our study as our problem aims to rank tags and comments of an image.

5.1 Methods

We evaluate the following 5 methods:

Random walk-based model (RW). A natural way to model and solve the tags and comments ranking problem is by leveraging a Markov random walk model, similar to the problem of tag ranking [7]. Given an image $(v, t, c)$, the comments $c$ and the tags $t$ and their relationships formulate a heterogeneous graph. Specifically, each comment $c \in c$ and each tag $t \in t$ is a node in the graph. Accordingly, there are three types of edges as follows.

Tag-tag similarity. Given a pair of tags (i.e., concepts) $t_i$ and $t_j$, the concept similarity between them, denoted as $\text{sim}(t_i, t_j) \in [0, 1]$, reflects both visual and semantic similarities of $t_i$ and $t_j$. The visual similarity measures the degree of visual similarity between the images annotated with tag $t_i$ and the images annotated with $t_j$. To this end, we adopt the exemplar similarity measure defined in [7]. For a tag $t_i$ of the given image, we select the $n$ nearest neighbors of images with tag $t_i$ to the image as exemplars of $t_i$, denoted by $N_i$. The exemplar similarity between tags $t_i$ and $t_j$ is $\text{sim}_v(t_i, t_j) = \exp \left( -\frac{n}{N_i} \sum_{x \in N_i, y \in N_j} d(x, y) \right)$ where $d(x, y) \in [0, 1]$ is the distance function measuring the visual distance between two images $x$ and $y$. In our experiments, we use cosine similarity of the low-level visual features to compute $d(x, y)$.

The semantic similarity is computed using tag co-occurrence as $\text{sim}_s(t_i, t_j) = f(t_i, t_j) / \sqrt{f(t_i) f(t_j)}$ where $f(x, y)$ is the frequency of $x$ and $y$ co-occurring in the same images and $f(x)$ is the tag frequency of $x$ in the whole collection. The concept similarity between tags $t_i$ and $t_j$ is a linear combination of the semantic and visual similarities: $\text{sim}(t_i, t_j) = \gamma \times \text{sim}_s(t_i, t_j) + (1 - \gamma) \times \text{sim}_v(t_i, t_j)$ where $\gamma \in [0, 1]$ controls the influence of visual similarity over semantic similarity. Here, $\text{sim}(t_i, t_j) = 1$ if $t_i$ and $t_j$ are identical tags and $\text{sim}(t_i, t_j) = 0$ if there is no semantic relationship between them. We set $\gamma = 0.8$.

Tag-Comment, and Comment-Comment Similarity. We define the tag-comment similarity between a tag $t_j$ and a comment $c_i$ to be the maximum similarity between $t_j$ and tags in $c_i$ i.e., $\text{sim}(c_i, t_j) = \max_{t_a \in c_i} \text{sim}(t_i, t_j)$. Given that each comment can be represented as a concept set (Section 3), we further extend the notion of concept similarity to compute comment-comment similarity: $\text{sim}(c_i, c_j) = \sum_{t_j \in c_j} \max_{t_a \in c_i} \text{sim}(t_a, t_j)$.

Because a tag $t_j$ can be considered as a concept set of size one, the tag-comment similarity is special case of comment-comment similarity. Note that $\text{sim}(c_i, c_j)$ may have a value larger than 1 if the two comments have more than one pair of highly similar tags.

Random Walk. The similarity between any two nodes depends on the types of the nodes. For easy presentation, we simply represent each node as a comment node because a tag node is equivalent to a comment node with a single concept. Considering the frequency of a concept appearing in tags and comments, the weight of the edge in the random walk graph between two nodes $c_i$ and $c_j$ is $\text{weight}(c_i, c_j) = \frac{1}{Z} \sum_{t_a \in c_i} \text{q}(t_a) \max_{t_j \in c_j} \text{q}(t_j) \text{sim}(t_a, t_j)$ where $\text{q}(x) = \text{count}(x) f(x)^{-1}$. Here $f(x)$ is the frequency of the concept in the whole dataset. $Z$ is the normalization factor to scale $\text{weight}(c_i, c_j)$ to $[0, 1]$. We can then derive the matrix of transition
probability to model the random walk process, to determine the relative importance of tags and comments.

**Visual Signature-based model (VS).** The method proposed in Section 3. We set $\theta = 1.5$, $\beta = 1$, $\delta = 0.05$, and $\lambda = 0.5$.

**The combined method (RW+VS).** It uses a simple voting strategy that averages the scores from the above two methods.

**Order-based Baseline (BL).** This method ranks tags and comments using the original chronological order.

**Tag Ranking (TA).** We use the method in [7]. Note that in [7], the tags are filtered by using Wikipedia. For a fair comparison with other methods in our evaluation, no tag filtering was conducted.

### 5.2 Results

**User Study.** We employ 12 human raters who assess the relevance of the ranked comments and tags in accordance to HCI research that recommends at least 10 users [5]. The experiments were run for 7 days. Each day 50 new images were randomly selected for evaluation and each selected image has at least 5 tags and at least one comment. In total 350 images were studied by each of the 12 volunteers. For each selected image, we presented it to a human rater along with top-5 ranked comments and tags generated using the different ranking methods. The rater was requested to compare and rate the comments and tags of each method with a score of 1 (irrelevant) to 5 (most relevant). We also asked the rater to rank methods from best (score is 1) to worst (score is 5). Figure 2(a) shows the average rank, given to a method in comparison with other methods, and average rating, given by the raters to the quality of the ranked comments/tags by the method. Hence, a model is superior if it has low rank and high rating scores.

We observe that methods that utilize both comments and tags significantly outperform the BL and TA approaches. In contrast, the differences in performance between VS and RW are relatively muted. Just utilizing the additional information provided by comments can improve the result quality regardless of the chosen technique. However, we note a slight improvement in average rating and rank using RW+VS, suggesting that any weaknesses inherent in either method could be alleviated through this combined strategy. Interestingly, the tag-based method (TA) could not perform better than the order-based baseline (BL). In our study, we utilized the full spectrum of tags (without pruning using WordNet or Wikipedia, for example). The added noise and complexity of the tag information resulted in existing tag-based methods being unable to outperform the order-based baseline. Note that BL itself is informative, because important concepts are likely to be created first.

**Effects of Number of Comments and Tags.** To understand the contribution of comments/tags to result quality, in this experiment we partition the images selected in the user study into different groups by their number of comments/tags.

Figure 2(b) plots the average user rating of the images within each group. We observe that when there are no more than 10 comments in an image, the performance of BL and TA become relatively closer to our three proposed methods. While methods that utilize both comments and tags remain superior to the ones that only use tags or chronological ordering, the gap is reduced compared to the images with more comments. For all other groups with more than 10 comments, however, we observe a clear improvement of our proposed methods over TA and BL. This suggests that with sufficient comments in the images, the prediction quality of relevant image tags become significantly superior when user comments are utilized. This reinforces our claim that user comments can be utilized to identify key concepts associated with images that attract users’ attention and interest. Figure 2(c) reports the impact of number of tags. We observe that the performance of our proposed methods remain largely unaffected between different groups. Across all groups, methods that incorporate comments outperform the BL and TA methods. This demonstrates that there is significant advantage in incorporatng user comments to identify interesting tags.

### 6 CONCLUSIONS

We have proposed a novel lightweight technique to concurrently identify and rank tags and comments associated with a social image that are relevant and have high user interest. Specifically, we introduce a visual signature-based model to find subsets of relevant comments and tags of a social image. Our user study demonstrated that utilization of both comments and tags to identify relevant tags significantly outperform techniques that rely solely on tags.

**REFERENCES**


