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

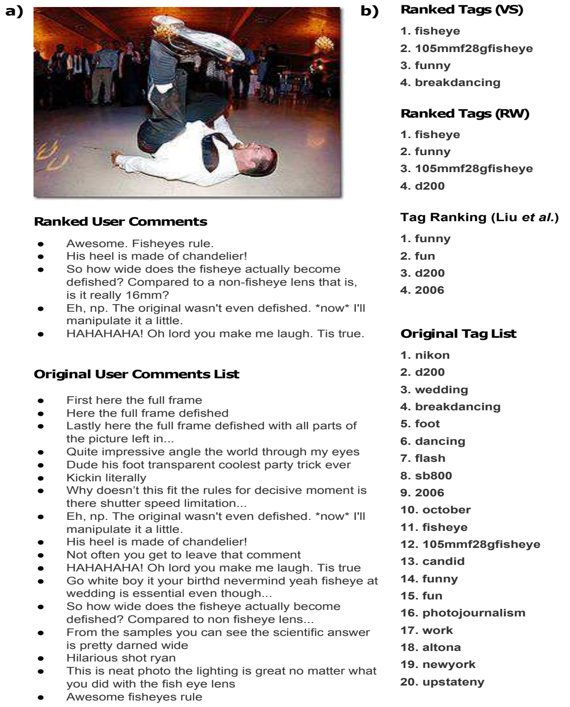
User-generated comments and tags can reveal important visual *concepts* associated with an image in Flickr. However, due to the inherent noisiness of the metadata, not all user tags are necessarily descriptive of the image. Likewise, comments may contain spam or chatter that are irrelevant to the image. 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:
Boon-Siew Seah   Aixin Sun   Sourav S Bhowmick. 2018. Killing Two Birds With One Stone: Concurrent Ranking of Tags and Comments of Social Images. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210095

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`).



Ranked User Comments

- Awesome. Fisheyes rule.
- His heel is made of chandelier!
- So how wide does the fisheye actually become defished? Compared to a non-fisheye lens that is, is it really 16mm?
- Eh, np. The original wasn't even defished. *now* I'll manipulate it a little.
- HAHAHAHA! Oh lord you make me laugh. Tis true.

Original User Comments List

- First here the full frame
- Here the full frame defished
- Lastly here the full frame defished with all parts of the picture left in...
- Quite impressive angle the world through my eyes
- Dude his foot transparent coolest party trick ever
- Kickin literally
- Why doesn't this fit the rules for decisive moment is there shutter speed limitation...
- Eh, np. The original wasn't even defished. *now* I'll manipulate it a little.
- His heel is made of chandelier!
- Not often you get to leave that comment
- HAHAHAHA! Oh lord you make me laugh. Tis true
- Go white boy it your birthd nevermind yeah fisheye at wedding is essential even though...
- So how wide does the fisheye actually become defished? Compared to non fisheye lens...
- From the samples you can see the scientific answer is pretty darned wide
- Hilarious shot ryan
- This is neat photo the lighting is great no matter what you did with the fish eye lens
- Awesome fisheyes rule

Ranked Tags (VS)

1. fisheye
2. 105mmf28gfisheye
3. funny
4. breakdancing

Ranked Tags (RW)

1. fisheye
2. funny
3. 105mmf28gfisheye
4. d200

Tag Ranking (Liu *et al.*)

1. funny
2. fun
3. d200
4. 2006

Original Tag List

1. nikon
2. d200
3. wedding
4. breakdancing
5. foot
6. dancing
7. flash
8. sb800
9. 2006
10. october
11. fisheye
12. 105mmf28gfisheye
13. candid
14. funny
15. fun
16. photojournalism
17. work
18. altona
19. newyork
20. upstateny

Figure 1: Illustration of comment and tag ranking.

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?

[1] The comments are collected separately through photo-ids.

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210095

(b) Which are the representative comments from users' discussion reflecting these concepts? Here, a *visual concept* refers to a concrete visual object or scene (*e.g.,* `cat`, `beach`), a visual effect that perceived by many users (*e.g.,* `fisheye`, `macro`), or an emotional effect arise from viewing the image (*e.g.,* `funny`, `scary`).

State-of-the-art tag ranking method rely on the visual and semantic similarities between tags to deduce the ranking among tags [7, 13]. In recent times, deep learning techniques have been employed for tag ranking and recommendation [3, 10]. However, all these techniques rank tags without leveraging the rich information hidden in users' comments. The goal of our research is to *concurrently* rank comments and tags associated with a social image, paving the way to identify most relevant comments *and* tags associated with 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.* in [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 $\langle \mathbf{v}, \mathbf{t}, \mathbf{c} \rangle$. The visual content of an image is represented by a set of visual features $\mathbf{v}$. Users may add comments about an image. By abusing the notation of lists, we represent these comments by the list $\mathbf{c} = [c_1, c_2, \ldots, c_m]$ where the comments are ordered by their time of posting. Each comment $c_i \in \mathbf{c}$ is modeled as a bag-of-words: $c_i = \{w_1, w_2, \ldots\}$. We denote the *frequency* of a word w appearing in a comment c_i as $count(w, c_i)$. The set of tags associated with an image is denoted as $\mathbf{t} = \{t_1, t_2, \ldots, t_n\}$. 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_i is associated with a *concept set*, denoted by $concept(c_i) = \{t_1, t_2, \ldots\}$. 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 $concept(c_i) = \emptyset$. Given a social image $\langle \mathbf{v}, \mathbf{t}, \mathbf{c} \rangle$, 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 $\langle \mathbf{v}, \mathbf{t}, \mathbf{c} \rangle$ and a word w, let $\mathbf{v}^{\mathbf{w}}$ be a vector of weights for the visual feature vector $\mathbf{v}$. 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* θ. If v_i is a significant visual feature of w, then $v_i^w > 0$; otherwise $v_i^w = 0$.

Visual Information Representation. We now extend the idea to represent an image $\langle \mathbf{v}, \mathbf{t}, \mathbf{c} \rangle$ via a subset of its comments $\mathbf{c}$ and tags $\mathbf{t}$. Given tags $\mathbf{t}$, the *visual information* of the image supported by the tags is defined as: $\mathbf{y_t} = \mathbf{v} \odot \left(\frac{1}{|\mathbf{t}|} \sum_{x \in \mathbf{t}} v^x \right)$ where $\odot$ is the entrywise product operation. Here $\mathbf{y_t}$ represents the visual information of the image that can be represented by the tags $\mathbf{t}$. Likewise, the *visual information* described by the entire corpus of comments is defined as $\mathbf{y_c} = \mathbf{v} \odot \left(\frac{1}{Z} \sum_{c_i} \sum_{x \in c_i} v^x count(x) \right)$ where $Z = \sum \sum_x count(x)$ normalizes the vector. Here $count(x)$ 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 $\mathbf{c_I} \subset \mathbf{c}$ such that $\mathbf{y_{c_I}}$ is sufficiently similar to the tags visual representation vector $\mathbf{y_t}$. (b) A subset of tags $\mathbf{t_I} \subset \mathbf{t}$ such that $\mathbf{y_{t_I}}$ is sufficiently similar to comments visual representation vector $\mathbf{y_c}$.

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 $\mathbf{c}$. The weight vector for selecting representative comments is represented by $\mathbf{w_c}$, and only positively weighted comments are selected. At the same time, the weight vector $\mathbf{w_t}$ selects representative tags. Then, we find the appropriate weights $\mathbf{w_c}$ and $\mathbf{w_t}$ by solving the following optimization problem:

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

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* (V_t, E_t) where V_t are the tag nodes and we add an edge $(i, j) \in E_t$ if $sim(i, j)$ is greater than a cut-off threshold δ. Given the graph, the *weight difference penalty function* is given by:

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

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

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

where λ specifies the penalty effect of the constraint graphs and β specifies the sparsity penalty. With the weight vectors $\mathbf{w_c}, \mathbf{w_t}$, we obtain the set of interesting comments and tags by choosing x whenever $\mathbf{w}(x) > 0$. 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 their weight values $\mathbf{w}(x)$.

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 $\langle \mathbf{v}, \mathbf{t}, \mathbf{c} \rangle$, the comments $\mathbf{c}$ and the tags $\mathbf{t}$ and their relationships formulate a heterogenous graph. Specifically, each comment $c \in \mathbf{c}$ and each tag $t \in \mathbf{t}$ is a node in the graph. Accordingly, there are three types of edges as follows.

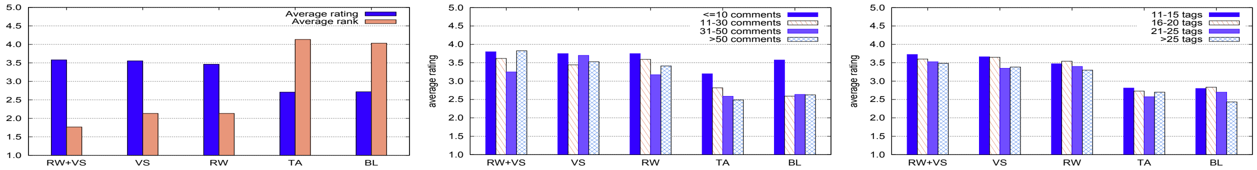
Tag-tag similarity. Given a pair of tags (*i.e.,* concepts) t_i and t_j, the *concept similarity* between them, denoted as $sim(t_i, t_j) \in [0, 1]$, reflects both *visual* and *semantic* similarities of t_i and t_j. The *visual similarity* measures the degree of visual similarity between the images 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 $sim_v(t_i, t_j) = exp\left(-\frac{1}{n^2}\sum_{x \in N_i, y \in N_j} d(x, y)\right)$ 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 $sim_s(t_i, t_j) = \frac{f(t_i, t_j)}{\sqrt{f(t_i)f(t_j)}}$ 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: $sim(t_i, t_j) = \gamma \times sim_s(t_i, t_j) + (1 - \gamma) \times sim_v(t_i, t_j)$ where $\gamma \in [0, 1]$ controls the influence of visual similarity over semantic similarity. Here, $sim(t_i, t_j) = 1$ if t_i and t_j are identical tags and $sim(t_i, t_j) = 0$ 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.,* $sim(c_i, t_j) = \max_{t_i \in c_i} sim(t_i, t_j)$. 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*: $sim(c_i, c_j) = \sum_{t_i \in c_i} \max_{t_j \in c_j} sim(t_i, t_j)$

Because a tag t_i can be considered as a concept set of size one, the tag-comment similarity is special case of comment-comment similarity. Note that $sim(c_i, c_j)$ 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 $weight(c_i, c_j) = \frac{1}{Z}\sum_{t_a \in c_i} q(t_a)\left(\max_{t_b \in c_j} q(t_b) sim(t_a, t_b)\right)$ where $q(x) = count(x)f(x)^{-1}$. Here $f(x)$ is the frequency of the concept in the whole dataset. Z is the normalization factor to scale $weight(c_i, c_j)$ to $[0, 1]$. We can then derive the matrix of transition



(a) Comparison of 5 methods (b) Evaluation by number of comments (c) Evaluation by number of tags

Figure 2: Comparative evaluation of the five methods.

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 incorporating 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

[1] Y.-Y. Chen, et al. Predicting Viewer Affective Comments Based on Image Content in Social Media. *In ICMR*, 2014.

[2] T.-S. Chua, et al. NUS-WIDE: a real-world web image database from National University of Singapore. *In CIVR*, 2009.

[3] S. Feng, Z. Feng, R. Jin. Learning to Rank Image Tags With Limited Training Examples. *IEEE TIP*, 24(4), 2015.

[4] Y. Gao, et al. Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search. *IEEE TIP* 22(1), 2013.

[5] J. Lazar, J. H. Feng, H. Hochheiser. Research Methods in Human-Computer Interaction. *John Wiley & Sons*, 2010.

[6] X. Li, C. G. M. Snoek, M. Worring. Learning Social Tag Relevance by Neighbor Voting. *IEEE TMM* 11(7), 2009.

[7] D. Liu, et al. Tag Ranking. *In WWW* 2009.

[8] E. Momeni, et al. Identification of useful user comments in social media. *In JCDL*, 2013.

[9] E. Momeni, et al. Leveraging Semantic Facets for Adaptive Ranking of Social Comments. *In ICMR*, 2017.

[10] H. T. H. Nguyen, et al. Personalized Tag Recommendation for Images Using Deep Transfer Learning. *In ECML/PKDD*, 2017.

[11] R. J. Tibshirani and J. Taylor The solution path of the generalized lasso. *The Annals of Statistics*, 39(3), 2011.

[12] Y. Wang, et al. Unsupervised Sentiment Analysis for Social Media Images. *In IJCAI*, 2015.

[13] L. Wu, Y. Wang, J. Shepherd. Efficient Image and Tag Co-ranking: A Bregman Divergence Optimization Method. *In ACM MM*, 2013.