10-2015

# Social tag relevance estimation via ranking-oriented neighbour voting

Chaoran CUI

Jialie SHEN
*Singapore Management University*, jlshen@smu.edu.sg

Jun MA

Tao LIAN

## Citation

CUI, Chaoran; SHEN, Jialie; MA, Jun; and LIAN, Tao. Social tag relevance estimation via ranking-oriented neighbour voting. (2015). *MM '15: Proceedings of the 23rd ACM International Conference on Multimedia: Brisbane, Australia, October 26-30, 2015*. 895-898.

**Available at:** https://ink.library.smu.edu.sg/sis_research/3541

MM '15: Proceedings of the 23rd ACM International Conference on Multimedia: Brisbane, Australia, October 26-30, 2015. pp. 895-898
http://doi.org/10.1145/2733373.2806358

# Social Tag Relevance Estimation via Ranking-Oriented Neighbour Voting

Chaoran Cui[1], Jialie Shen[2], Jun Ma[1], and Tao Lian[1]
[1] School of Computer Science and Technology, Shandong University, Jinan, China
[2] School of Information Systems, Singapore Management University, Singapore
{bruincui, liantao1988}@gmail.com, jlshen@smu.edu.sg, majun@sdu.edu.cn

## ABSTRACT

User-generated tags associated with social images are frequently imprecise and incomplete. Therefore, a fundamental challenge in tag-based applications is the problem of tag relevance estimation, which concerns how to interpret and quantify the relevance of a tag with respect to the contents of an image. In this paper, we address the key problem from a new perspective of learning to rank, and develop a novel approach to facilitate tag relevance estimation to directly optimize the ranking performance of tag-based image search. A supervision step is introduced into the neighbour voting scheme, in which tag relevance is estimated by accumulating votes from visual neighbours. Through explicitly modelling the neighbour weights and tag correlations, the risk of making heuristic assumptions is effectively avoided for conventional methods. Extensive experiments on a benchmark dataset in comparison with the state-of-the-art methods demonstrate the promise of our approach.

## Categories and Subject Descriptors

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

## Keywords

Tag-based image search; Tag relevance estimation; Learning to rank; Neighbour voting

## 1. INTRODUCTION

With the advent of Web 2.0 technology, we have witnessed an explosive growth of social images in recent years [1, 9]. This also raises an urgent demand for smart search technologies to explore large scale social image repositories, such as Flickr and Pinterest. Rather than simply providing the interfaces for image storage and sharing, these repositories allow users to label images with freely-chosen tags. Naturally, these tags can be used to index social images to facilitate the search process. As a result, tag-based image search has become the *de facto* way to access and browse resources in current social image repositories.

Despite the popularity of tag-based image search, its performance is still far from satisfactory due to the inferior quality of user-generated tags, which are often imprecise and incomplete for describing the image contents, and thus cannot be regarded as qualified indexing keywords for image search. Extensive research endeavours have been devoted to improving the descriptive capability of tags regarding the image contents [7, 8]. Although varying in terms of their targeted tasks and methodologies, these works all rely on the key functionality of tag relevance [6], i.e., estimating the relevance of a tag with respect to the visual content of an image, which is also referred to as the problem of *tag relevance estimation.*

Tag relevance estimation has been widely studied, however, there remains a common limitation in the literature. In essence, the aim of tag relevance estimation is to find out the tags that can be applied as content descriptors for the images, so that the accuracy of image search can be further improved by indexing the images with these reliable tags. However, existing studies generally perform tag relevance estimation without the explicit intention of promoting the performance of tag-based image search. In most cases, they estimate the tag relevance via heuristic rules [5], through optimizing the classification accuracy for specific tags [10], or by maximizing the likelihood of the annotations of training images [10]. We argue that these objectives are not directly related to the search performance, and optimizing them does not necessarily yield good search results.

In light of the above deficiency, in this paper, we investigate the problem from a new perspective of learning to rank, and perform tag relevance estimation to directly optimize the ranking performance of tag-based image search. Specifically, we introduce a supervision step into the neighbour voting scheme [5], in which the tag relevance is estimated by accumulating votes from the visual neighbours of the given image. Our approach explicitly models the neighbour weights and tag correlations, and thereby avoids the risk of making heuristic assumptions for conventional methods. Experimental results demonstrate the promise of our approach in both applications of tag-based image search and automatic tag recommendation.

## 2. TAG RELEVANCE FORMULATION

Our framework formulates a tag relevance function based on the neighbour voting scheme, where the relevance score of a tag is inferred by the tagging information of the visual neighbours of the given image. While the standard neighbour voting algorithm [5] simply let the neighbours vote equally, efforts have been made to weight neighbours in terms of their importance. Typically, the weights are heuristically determined by the visual similarity [4], or in a soft manner that performs a random walk over the $k$-nearest neighbor graph [13]. However, the performance gain obtained by these heuristic weighting strategies appears to be limited [13]. In this paper, we propose to explicitly model the individual weight of each visual neighbour. Given a tag $t_i$ and an image $x$, we define $r(t_i, x)$ as the relevance score of $t_i$ with respect to $x$:

$$r(t_i, x) = \sum_{l=1}^{k} v_l \varphi(d_l^x, t_i) \ , \tag{1}$$

where $d_l^x$ denotes the $l$-th nearest neighbour of $x$, $\mathbf{v} \in \mathbb{R}^{k \times 1}$ is a vector of parameters whose $l$-th element $v_l$ indicates the weight of $d_l^x$, and $k$ is the total number of the visual neighbours. Note that we treat an image itself as its first nearest neighbour, i.e., $d_1^x = x$. $\varphi(d_l^x, t_i)$ represents the confidence value that $t_i$ is relevant to $d_l^x$, which is estimated by a multiple Bernoulli process with a beta prior:

$$\varphi(d_l^x, t_i) = \frac{\mu \delta_{d_l^x, t_i} + s_{t_i}}{\mu + s} \ , \tag{2}$$

where $\delta_{d_l^x, t_i}$ indicates the tagging observation on $d_l^x$, i.e., $\delta_{d_l^x, t_i} = 1$ if $d_l^x$ is tagged with $t_i$ in the image collection and zero otherwise. $\mu$ is a smoothing parameter associated with $\delta_{d_l^x, t_i}$. $s_{t_i}$ denotes the number of images tagged with $t_i$, and $s$ is the total number of images.

It has proven that the information of tag correlations is highly beneficial to many tag-based applications [7]. Motivated by this, we seek to incorporate the tag correlations to facilitate tag relevance estimation. Co-occurrence statistics and WordNet similarity are the most commonly used correlation measurements. However, these prior information may be unreliable when dealing with noisy social tags. More importantly, limited by their non-negative property, both co-occurrence statistics and WordNet similarity cannot capture the potential negative correlations among tags. Therefore, we propose to explicitly model the pairwise tag correlations and to reformulate the relevance function as follows:

$$r(t_i, x) = w_{ii} \sum_{l=1}^{k} v_l \varphi(d_l^x, t_i) + \sum_{j=1, j \neq i}^{m} w_{ij} \sum_{l=1}^{k} v_l \varphi(d_l^x, t_j) \ , \tag{3}$$

where $W \in \mathbb{R}^{m \times m}$ is a parameter matrix whose $(i, j)$-th entry $w_{ij}$ represents the correlation between the tag $t_i$ and the tag $t_j$, and $m$ is the total number of unique tags. We assume that $W$ is a symmetric matrix, i.e., $w_{ij} = w_{ji}$, and both positive and negative values are allowed in $W$. In Equation (3), we exploit not only the confidence score of $t_i$ being relevant to the neighbours of $x$, but also the evidences provided by all the other tags. For the simplicity of the expression, we introduce a supplementary matrix $\Phi_x \in \mathbb{R}^{k \times m}$ whose $(l, j)$-th entry is equal to $\varphi(d_l^x, t_j)$. As a result, Equation (3) can be written in a concise form:

$$r(t_i, x) = \mathbf{e}_i{}^T W \Phi_x{}^T \mathbf{v} \ . \tag{4}$$

A potential problem with the above formulation is that it requires a large number of parameters to capture the correlation between each pair of tags. From the viewpoint of statistical learning theory, too many parameters may degrade the model stability and generalization in performance. In light of this, we further introduce a low-rank prior into the parameter $W$ with $W = U^T U$, which results in the new formulation of the relevance function as follows:

$$r(t_i, x) = \mathbf{e}_i{}^T U^T U \Phi_x{}^T \mathbf{v} \ , \tag{5}$$

where $U \in \mathbb{R}^{p \times m}$ and $p$ is the dimensionality of a latent space. Let $\mathbf{u}_i$ denote the $i$-th column of $U$, which corresponds to the representation vector of $t_i$ in the latent space. The correlation $w_{ij}$ is thus measured by the dot product of $\mathbf{u}_i$ and $\mathbf{u}_j$, which is commonly used to measure the matching between textual vectors. Because the intrinsic dimensionality of the latent space is typically much smaller than that of the original space (i.e., $p \ll m$), the number of parameters in Equation (5) is significantly reduced.

## 3. RANKING-ORIENTED LEARNING

Our framework seeks to learn the parameters $\mathbf{v}$ and $U$ in a supervised fashion, and to facilitate tag relevance estimation to boost the accuracy of image search. To this end, we consider the relevance function $r$ as a ranking criterion for tag-based image search, and approach the problem from a new perspective of learning to rank.

Without loss of generality, assume that a set of training instances for the first $n$ tags is available:

$$\{(t_q, Y_q^*) \ \in \mathcal{T} \times \mathcal{Y} : q = 1, \ldots, n\} \ ,$$

where $t_q$ is a query tag and $Y_q^*$ is the true ranking of the images with respect to $t_q$. $\mathcal{T}$ is the tag vocabulary and $\mathcal{Y}$ is the set of all possible rankings over images. Similar to [11], we represent any ranking $Y \in \mathcal{Y}$ as a matrix of pair orderings, whose $(i, j)$-th entry $y_{ij} = +1$ if the image $x_i$ is ranked ahead of the image $x_j$, $y_{ij} = -1$ if $x_i$ is ranked behind $x_j$, and $y_{ij} = 0$ if $x_i$ and $x_j$ have equal rank. Note that $Y_q^*$ is a weak ranking with only two relevance levels, i.e., relevant and irrelevant. We denote by $\mathcal{I}_{t_q}^+$ and $\mathcal{I}_{t_q}^-$ the sets of the relevant and irrelevant images with respect to $t_q$.

Our goal is transformed into learning a ranking hypothesis $h : \mathcal{T} \to \mathcal{Y}$. For the query tag $t_q$, $h(t_q)$ needs to correspond to the image ranking in descending order of $r(t_q, x)$. For this purpose, we first construct a compatibility function $f(t_q, Y)$ measuring how well a possible image ranking $Y$ fits for $t_q$:

$$\begin{aligned} f(t_q, Y) &= \sum_{x_i \in \mathcal{I}_{t_q}^+} \sum_{x_j \in \mathcal{I}_{t_q}^-} y_{ij} \left( \frac{r(t_q, x_i) - r(t_q, x_j)}{|\mathcal{I}_{t_q}^+| \cdot |\mathcal{I}_{t_q}^-|} \right) \qquad (6) \\ &= \sum_{x_i \in \mathcal{I}_{t_q}^+} \sum_{x_j \in \mathcal{I}_{t_q}^-} y_{ij} \left( \frac{\mathbf{e}_q{}^T U^T U (\Phi_{x_i} - \Phi_{x_j})^T \mathbf{v}}{|\mathcal{I}_{t_q}^+| \cdot |\mathcal{I}_{t_q}^-|} \right) . \end{aligned}$$

$h(t_q)$ is then defined by maximizing $f(t_q, Y)$ over $Y \in \mathcal{Y}$:

$$h(t_q) = \arg\max_{Y \in \mathcal{Y}} f(t_q, Y) \ . \tag{7}$$

Here, $f(t_q, Y)$ is decomposed into a series of pairwise components, i.e., $y_{ij}(r(t_q, x_i) - r(t_q, x_j))$. For the fixed $\mathbf{v}$ and $U$, $h(t_q)$ can be attained by maximizing each component individually: if $r(t_q, x_i) > r(t_q, x_j)$, $y_{ij}$ is set to $+1$; otherwise, it is set to $-1$. Note that this is the same procedure as sorting the images by $r(t_q, x)$, and $h(t_q)$ proves to be equivalent to the ranking in descending order of $r(t_q, x)$.

With the set of training instances, we learn the ranking hypothesis $h(t_q)$ by minimizing the empirical ranking risk,

$$R_\Delta(h) = \frac{1}{n} \sum_{q=1}^{n} \Delta(Y_q^*, h(t_q)) \ , \tag{8}$$

where the loss function $\Delta$ quantifies the inconsistency between the derived ranking $h(t_q)$ and the true ranking $Y_q^*$. In our study, we define $\Delta$ based on AP score:

$$\Delta(Y_q^*, h(t_q)) = 1 - AP(Y_q^*, h(t_q)) \ . \tag{9}$$

As a result, minimizing $R_\Delta(h)$ is equivalent to directly optimizing MAP performance of tag-based image search.

We adopt the structural SVM [3] as the backbone of our learning algorithm. Specifically, we rewrite the compatibility function $f$ as:

$$f(t_q, Y) = \left\langle U^T U \otimes \mathbf{v},\ \mathbf{e}_q \otimes \Psi(t_q, Y) \right\rangle_F \ , \tag{10}$$

$$\Psi(t_q, Y) = \sum_{x_i \in \mathcal{I}_{t_q}^+} \sum_{x_j \in \mathcal{I}_{t_q}^-} y_{ij} \frac{\Phi_{x_i} - \Phi_{x_j}}{|\mathcal{I}_{t_q}^+| \cdot |\mathcal{I}_{t_q}^-|} \ , \tag{11}$$

where $\otimes$ denotes the Kronecker product, and $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. $\Psi(t_q, Y)$ encodes the *joint features* of the input-output pair ($t_q$, $Y$). By representing $f$ in such a form of a linear function of $\Psi(t_q, Y)$, the structural SVM can be employed to learn $\mathbf{v}$ and $U$ through the following optimization problem [3]:

Optimization Problem 1.

$$\underset{V,U,\xi}{min.} \quad \frac{\lambda}{2} \|V\|_2^2 + \frac{\lambda}{2} \|U\|_F^2 + \xi \tag{12}$$

$$s.t. \qquad \forall (Y_1, \ldots, Y_n) \in \mathcal{Y}^n :$$

$$\frac{1}{n} \sum_{q=1}^{n} \left[ f(t_q, Y_q^*) - f(t_q, Y_q) \right] \geqslant \frac{1}{n} \sum_{q=1}^{n} \Delta(Y_q^*, Y_q) - \xi \ . \tag{13}$$

The main difficulty of Optimization Problem 1 is that there are as many as $|\mathcal{Y}|^n$ constraints to be considered. To solve it efficiently, we employ the cutting plane algorithm. The pseudo-code of the algorithm is presented in Algorithm 1. It starts with an empty working set $\mathcal{W}$ of active constraints (step 1). Then it iteratively finds the $\widehat{Y_q}$ which generates the most violated constraint for each $t_q$ with current $\mathbf{v}$ and $U$ (step 3-5). For our MAP-related loss, this step can be accomplished with the method presented in [11]. If the corresponding constraints are on average violated by more than the error tolerance $\varepsilon$, the algorithm adds $(\widehat{Y}_1, \ldots, \widehat{Y}_n)$ into $\mathcal{W}$, and re-optimizes over the updated $\mathcal{W}$ (step 6-9). We implement a sub-gradient method to solve the problem at this step. The sub-gradients with respect to $\mathbf{v}$ and $U$ can be computed in terms of the single constraint $(\widehat{Y}_1, \ldots, \widehat{Y}_n)$ that achieves the current largest margin in $\mathcal{W}$:

$$\nabla_{\mathbf{v}} = \lambda \mathbf{v} - \frac{1}{n} \sum_{q=1}^{n} \delta\Psi(t_q, \widehat{Y}_q) U^T U \mathbf{e}_q$$

$$\nabla_{\mathrm{U}} = \lambda U - \frac{1}{n} \sum_{q=1}^{n} \Big( \mathbf{v}^T \delta\Psi(t_q, \widehat{Y}_q) \otimes \mathbf{u}_q \ , \tag{14}$$

$$+ {\mathbf{e}_q}^T \otimes U \delta\Psi(t_q, \widehat{Y}_q)^T \mathbf{v} \Big)$$

where $\delta\Psi(t_q, \widehat{Y}_q) = \Psi(t_q, Y_q^*) - \Psi(t_q, \widehat{Y}_q)$. Algorithm 1 terminates when no constraints are added into $\mathcal{W}$ (step 10).

# 4. EXPERIMENTS

## 4.1 Data Collections

All the experiments are conducted on a benchmark image dataset collected from Flickr, i.e., NUS-WIDE-Lite (NUS for short hereafter) [2]. On NUS, the labelling ground-truth of the images for 75 *concepts* has been provided. These concepts are used as the query tags in the experiments.

**Algorithm 1** Cutting Plane Algorithm

```
Input: training instances $(t_1, Y_1^*), \ldots, (t_n, Y_n^*)$, regularization trade-off $\lambda$, error tolerance $\varepsilon$
Output: model parameters $\mathbf{v}$ and $U$, slack variable $\xi$
 1: Initialize $\mathcal{W} \leftarrow \emptyset$
 2: repeat
 3:    for $q = 1, 2, \ldots, n$ do
 4:       $\widehat{Y}_q \leftarrow \arg\max_{Y \in \mathcal{Y}} \Delta(Y_q^*, Y) + f(t_q, Y)$
 5:    end for
 6:    if $\frac{1}{n} \sum_{q=1}^{n} \Delta(Y_q^*, \widehat{Y}_q) - \frac{1}{n} \sum_{q=1}^{n} \left[ f(t_q, Y_q^*) - f(t_q, \widehat{Y}_q) \right] > \xi + \varepsilon$
       then
 7:       $\mathcal{W} \leftarrow \mathcal{W} \cup \left\{ (\widehat{Y}_1, \ldots, \widehat{Y}_n) \right\}$
 8:       Optimize Equation (12) over $\mathcal{W}$
 9:    end if
10: until $\mathcal{W}$ has no change during iteration
```

Each image is represented with the same features as described in [2]. We use the Euclidean metric to measure the visual distance between images. Given an image, all images are ranked by their distance from it and the $k$ nearest neighbours are subsequently discovered.

## 4.2 Evaluation Methodology

We compare our approach with several state-of-the-art methods on tag relevance estimation. Specifically, the competitors include: NVote [5], WVote [4], Graph [12], SVM [10] and TagProp [10]. We denote our approach as RankVote.

To conduct a comprehensive comparison, we first evaluate in the scenario of tag-based image search. Given a query tag, we sort images by descending the predicted tag relevance concerning each image. We study the performance of different methods with regard to the parameter $k$. We take half of the total concepts as query tags during training, and keep the rest for testing. We use MAP as the evaluation metric.

We also compare different methods in the scenario of automatic tag recommendation. Note that SVM is excluded from this evaluation because of its excessive computational cost in the need of learning a separate classifier for each tag. We randomly choose 1,000 images as the evaluation testbed. Given an image, all tags are ranked by their relevance values, and the top 10 results are added into the recommendation list. Three volunteers manually label each recommended tag with four relevance levels: Most Relevant (score 3), Relevant (score 2), Weakly Relevant (score 1) and Irrelevant (score 0). As the ground-truth tag ranking for a single image is not available, DCG is used to evaluate the quality of the recommendation list.

## 4.3 Experimental Results

### 4.3.1 Performance Comparison

Figure 1 shows the image search results of different methods. Clearly, RankVote consistently achieves the best performance in all cases, reaching at least 20% relative improvement over the other methods. Besides, RankVote is also more robust to the variation of $k$. Even though tested with only 10 neighbours, RankVote still remains a relatively high performance level of 49.9%. The results verify the promise of RankVote for tag-based image search.

Table 1 reports the empirical results for automatic tag recommendation. It is shown that RankVote outperforms

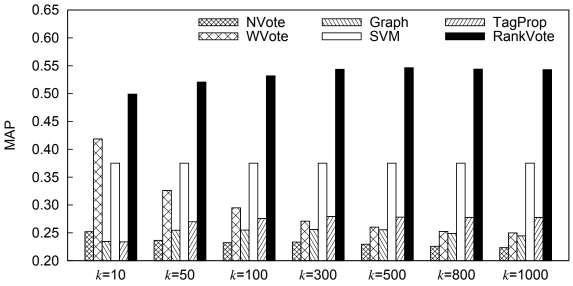

**Figure 1: Performance comparison for image search.**

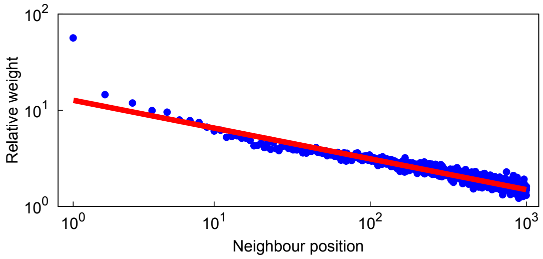

**Figure 2: Illustration of neighbour weights.**

the other approaches in both evaluation criteria. Especially, we find that the relative improvements on DCG@5 are more significant than those on DCG@10. It is a nice property as users are usually more interested in the top results in recommendation. The conclusion can thus be drawn that RankVote is an effective technique for tag recommendation.

### *4.3.2 Illustration of Learned Parameters*

To the best of our knowledge, this is the first study that effectively estimates the tag relevance, while jointly learns the neighbour weights and tag correlations in a unified framework. Figure 2 plots the individual weight of each neighbour in the top 1,000 sequence on a log-log scale. We notice that the variation trend of neighbour weights is in accordance with the intuition that close neighbours are more important than those distant ones, and it can also be well fitted by a power-law relationship, which has been indicated by a red line. We believe this finding is valuable to guide the future research on neighbour voting.

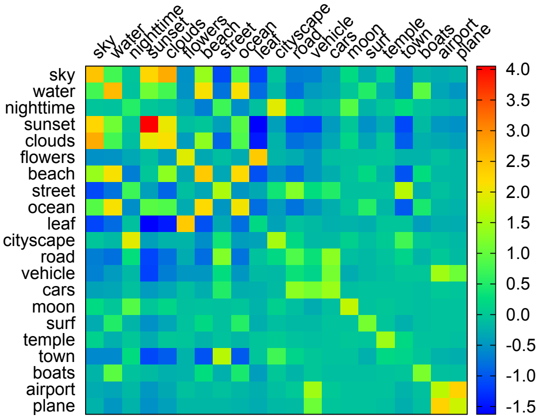

**Figure 3: Illustration of tag correlations.**

**Table 1: Results for automatic tag recommendation.**

| | NVote | WVote | Graph | TagProp | RankVote |
|---|---|---|---|---|---|
| DCG@5 | 3.50 | 3.29 | 2.46 | 3.65 | **4.43** |
| DCG@10 | 4.82 | 4.49 | 3.57 | 4.88 | **5.72** |

Figure 3 illustrates the learned pairwise correlations among a group of frequent tags, where a colour map is used to indicate the magnitude of tag correlations. Among the pairwise elements, the higher correlations are assigned to the pairs of tags that commonly co-occur such as (sky, clouds), or the tags with the same or similar meanings such as (water, ocean). On the other hand, those rarely co-occurring tags like (sunset, leaf) and (ocean, town) are assigned with lower negative correlation values. This reveals that the learned correlations can properly encode the various kinds of relationships among tags.

## 5. CONCLUSIONS

In this paper, we investigate the problem of tag relevance estimation from a new perspective of learning to rank. A supervision step is introduced into the neighbour voting scheme, and both neighbour weights and tag correlations are explicitly modelled. Experimental results demonstrate the effectiveness of our approach in both scenarios of image search and tag recommendation. For future study, we plan to incorporate the distance metric learning techniques for discovering more accurate visual neighbours.

## 6. ACKNOWLEDGEMENTS

This work is supported by the Natural Science Foundation of China (61272240, 71402083, 61103151), the Doctoral Fund of Ministry of Education of China (20110131110028) and the Natural Science Foundation of Shandong province (ZR2012FM037). Dr. Jialie Shen was supported by Singapore Ministry of Education under Academic Research Fund Tier-2 (MOE Ref: MOE2013-T2-2-156) and the Microsoft Research Faculty Grant (FY14-RES-OPP-048).

## 7. REFERENCES

[1] Z. Cheng, J. Shen, and H. Miao. The effects of multiple query evidences on social image retrieval. *Multimedia Systems*, 2014.

[2] T.-S. Chua and J. Tang. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, 2009.

[3] T. Joachims, T. Finley, and C. Yu. Cutting-plane training of structural svms. *Machine Learning*, 2009.

[4] S. Lee and W. De Neve. Visually weighted neighbor voting for image tag relevance learning. *Multimed Tools Appl.*, 2014.

[5] X. Li, C. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *TMM*, 2009.

[6] X. Li, T. Uricchio, and L. Ballan. Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval. *arXiv preprint arXiv:1503.08248*, 2015.

[7] D. Liu, X. Hua, and H. Zhang. Content-based tag processing for internet social images. *Multimed Tools Appl.*, 2011.

[8] J. Shen, M. Wang, and T.-S. Chua. Accurate online video tagging via probabilistic hybrid modeling. *Multimedia Systems*, 2014.

[9] J. Shen, M. Wang, S. Yan, and X.-S. Hua. Multimedia tagging: past, present and future. In *MM*, 2011.

[10] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid. Image annotation with tagprop on the mirflickr set. In *MIR*, 2010.

[11] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *SIGIR*, 2007.

[12] D. Zhou, O. Bousquet, and T. Lal. Learning with local and global consistency. In *NIPS*, 2004.

[13] X. Zhu and W. Nejdl. An adaptive teleportation random walk model for learning social tag relevance. In *SIGIR*, 2014.