Content is Still King: The Effect of Neighbor Voting Schemes on Tag Relevance for Social Image Retrieval[*]

Ba Quan Truong
bqtruong@ntu.edu.sg

Aixin Sun
axsun@ntu.edu.sg

Sourav S. Bhowmick
assourav@ntu.edu.sg

School of Computer Engineering, Nanyang Technological University, Singapore 639798

ABSTRACT

Tags associated with social images are valuable information source for superior tag-based image retrieval (TagIR) experiences. One of the key issues in TagIR is to learn the effectiveness of a tag in describing the visual content of its annotated image, also known as *tag relevance*. One of the most effective approaches in the literature for tag relevance learning is *neighbor voting*. In this approach a tag is considered more relevant to its annotated image (also known as the *seed* image) if the tag is also used to annotate the *neighbor* images (nearest neighbors by visual similarity). However, the state-of-the-art approach that realizes the neighbor voting scheme does not explore the possibility of exploiting the *content* (*e.g.,* degree of visual similarity between the seed and neighbor images) and *contextual* (*e.g.,* tag association by co-occurrence) features of social images to further boost the accuracy of TagIR. In this paper, we identify and explore the viability of four *content* and *context-based dimensions* namely, *image similarity*, *tag matching*, *tag influence*, and *refined tag relevance*, in the context of tag relevance learning for TagIR. With alternative formulations under each dimension, this paper empirically evaluated 20 neighbor voting schemes with 81 single-tag queries on nus-wide dataset. Despite the potential benefits that the contextual information related to tags bring in to image search, surprisingly, our experimental results reveal that the content-based (image similarity) dimension is still the king as it significantly improves the accuracy of tag relevance learning for TagIR. On the other hand, tag relevance learning does not benefit from the context-based dimensions in the voting schemes.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation*

Keywords

Social Image, Flickr, Tag Relevance, Neighbor Voting

1. INTRODUCTION

The increasing prevalence of digital photography devices (*e.g.,* digital cameras, camera-enabled mobile phones) and increasing popularity of social image sharing platforms (*e.g.,* `Flickr`) have made availability of huge volume of images online. At the same time, finding relevant images that best match a particular user's information need (*i.e.,* the task of image retrieval) has become extremely daunting. This has led to increasing research efforts by academic and commercial communities toward building superior image retrieval strategies. Specifically, image retrieval has been widely studied from two paradigms, namely *content-based* and *annotation-based* image retrieval [2, 4]. The former relies on visual descriptors extracted from the images and aims to return images that best match a user-specified example image. The latter returns images matching user's keyword query mainly based on the keyword annotations assigned to images. While obtaining high-quality image annotations, either manually or automatically, has always been a major obstacle for annotation-based image retrieval, freely available tags in social image sharing platforms have become a valuable alternative source of annotations. Consequently, *tag-based image retrieval* (TagIR), which is the task of retrieving the best matching images for a keyword query based on the tags assigned to the images, has enjoyed increasing attention by the research community and end users.

In a recent study related to different TagIR methods, Sun et al. [18] identified five dimensions to quantify the *matching score* between a tagged image and a keyword query and empirically evaluated the impact of these dimensions. A key finding in this study is that the degree of effectiveness of a tag in describing the tagged image (known as *tag relatedness* or *tag relevance*) is one of the most crucial dimensions for superior TagIR experience, especially for single-tag queries. In fact, the best performing tag relevance measure used in [18] is based on the notion of *neighbor voting* as proposed by Li et al [7,8]. The idea is that if many distinct users use the same tags to label visually similar images, then these tags are likely to reflect the visual contents of the annotated images. Specifically, given an image d (*seed image*), its k-nearest neighbors (*neighbor images*) are first obtained based on some visual similarity measure. Then, the tag relevance of a tag $t \in d$ is the probability of t being used to annotate the neighborhood images minus the probability of the tag being used in the entire collection, *i.e.,* $P(t|N(d)) - P(t|\mathcal{D})$, where $N(d)$ is the set of visually nearest neighbors of image d and $\mathcal{D}$ denotes the image collection (see Section 2 for more details). In this paper, this simple voting scheme is referred to as the *baseline scheme*.

Although the aforementioned scheme is simple and effective, we observe that it is oblivious to several distinct *content* and *contextual* characteristics of social images, which if exploited may further en-

[*] This work was supported by Singapore MOE AcRF Tier-1 Grant RG13/10.

ICMR '12, June 5-8, Hong Kong, China
Copyright ©2012 ACM 978-1-4503-1329-2/12/06 ...$10.00.

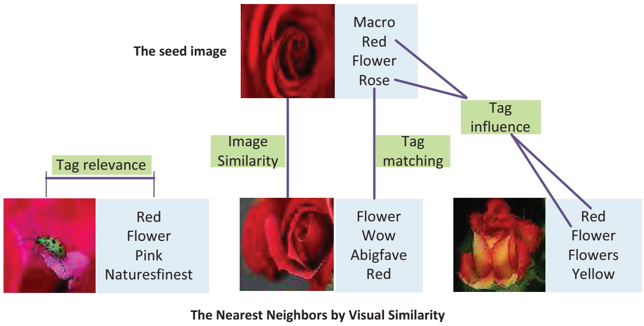


Figure 1: Dimensions of tag relevance by neighbor voting.

hance the accuracy of tag-based image retrieval. Let us elaborate on this further. Consider the example of neighbor voting in Figure 1. According to the baseline scheme, every neighbor image contributes equally to the voting *without* considering the similarity between the neighbor images and the seed image. However, it is clear that the second neighbor image is visually more similar to the seed image and should be allocated more voting power. Such visual content information is not exploited by the baseline scheme.

There are also several distinct *contextual* characteristics due to the nature of social tagging that may provide us opportunities to boost the accuracy of TagIR. Firstly, social tags are assigned by common users and are noisy in nature. Users may not use exactly the same tag to describe a similar visual concept. For instance, the flower tag in the second neighbor image and the rose tag in the seed image carry similar meaning. Note that in the baseline scheme if a neighbor image has a tag that exactly matches a tag in the seed image, then it has one vote towards that tag, otherwise zero. Secondly, users may often use multiple tags instead of a single tag to describe the visual content of the image. For instance, the tags red and rose are used to describe the seed image. Similarly, red and flower are used to describe all the neighbor images. Observe that both red and rose (resp. flower) are more relevant to the seed (resp. neighbor) image compared to either red or rose (resp. flower). The baseline scheme, however, assumes that the visual content of an image is completely described by only single tags. Thirdly, not all tags of an image (both seed and neighbor) equally describe the visual content of the image. For instance, flower is more effective in describing the content of the first neighbor image compared to the tag Naturesfinest. Although the baseline scheme considers the relative relevance of the tags in the seed image for tag relevance learning, it treats the relevance of all tags of a neighbor image equally and gives them the same voting power.

Is it possible to enhance TagIR accuracy by incorporating the aforementioned content and contextual features of social images in tag relevance learning? In this paper, we show how to incorporate these features in neighbor voting-based tag relevance computation to address this question. Specifically, we incorporate the following four dimensions (one *content-based* and three *context-based*) to learn tag relevance.

1. **Image Similarity**. This content-based dimension allocates more voting power to those neighbor images that are more visually similar to the seed image.

2. **Tag Matching**. Since tags are incomplete, this dimension incorporates *tag association measures* to allocate some amount of voting power to a neighbor image even if it does not contain a tag of the seed image but contains some tags that are highly associated with it.

3. **Tag Influence**. This dimension explores the relevance of multiple tags for jointly describing the visual content of an image.

4. **Refined Tag Relevance**. In this dimension, tag relevance is learned for *every* image (instead of just the seed image) in the collection and then the learned tag relevance is used to further refine tag relevance learning of the seed image by giving different voting power.

To gain an in-depth understanding of the impact of the above dimensions, we systematically evaluated the alternative formulations under each dimension and their combinations for tag relevance voting. As tag relevance is difficult to be evaluated directly, we adopt an extrinsic evaluation approach to evaluate the learned tag relevance using the TagIR task based on the settings described in [18]. The evaluation was conducted on nus-wide dataset [3], the largest human-annotated dataset consisting of more than 269K images from `Flickr`, and involves 20 voting schemes in total with 81 single-tag queries.

We use the *Precision at top-K* (*P@K*) and *Mean Average Precision* (map) measures to report a detailed analysis on the performance of different voting schemes and the impact of the formulations under each dimension. Although at first glance it may seem that the three context-based dimensions (*e.g.,* tag matching, tag influence, refined tag relevance) should enhance the accuracy of TagIR, *our experimental results suggest that the content-based dimension (image similarity) instead of these context-based ones is still the king!* That is, the visual similarity between the seed image and neighbor images significantly improves TagIR accuracy. Interestingly, voting schemes involving tag associations through formulations under tag matching and tag influence models generally affect the TagIR accuracy adversely. Surprisingly, our results also show that using the learned tag relevance in the image collection to refine tag relevance learning does not benefit most queries, leading to poorer overall TagIR accuracy. However, we note that these negative results should not be interpreted as that contextual features (*e.g.,* tag co-occurrence) are not beneficial in general in tag-based social image retrieval. Tag co-occurrence does benefit retrieval accuracy if considered under some other dimensions (*e.g.,* tag-query matching model) of TagIR [18], but does not in learning tag relevance for a given seed image from its visually similar neighbors.

The rest of the paper is organized as follows. We review related research in Section 2. In Section 3, we present the content and context-based dimensions and formulations for tag relevance voting schemes. In Section 4, we report the experimental setup for the evaluation of the 20 voting schemes with alternative formulations for the four dimensions. Section 5 presents systematic analysis of the experimental results on the impact of the dimensions in the voting schemes for TagIR task followed by a discussion on our interpretations of the key experimental findings. The last section concludes the paper.

2. RELATED WORK

Social Image Tag Relevance Learning. Social image tagging has gained significant attention across a wide spectrum of dimensions from various research communities due to its potential to support superior image retrieval and related applications. These dimensions include the study on motivations for social image tagging [1], taxonomy and comparison of tagging systems [13], tag types [15, 16]. Since tags are noisy and incomplete in nature, several recent research on social image tagging have focused on tag recommendation, disambiguation, and de-noising among others [3]. In the following, we review the works related to tag relevance learning *i.e.,* determining the effectiveness of a tag in describing the visual content of the tagged image.

Li *et al.* in [8] proposed to learn tag relevance by visual nearest neighbor voting based on two assumptions known as *user tagging* and *visual search*. The former assumes that the probability of correct user tagging is larger than the probability of incorrect tagging in a large user-tagged image collection; the latter assumes a content-based visual search is better than random sampling. For a given image d its tag relevances are then computed in two steps. In the first step, it obtains the K nearest neighbors of d based on the visual features of images under *unique user constraint* (a user has at most one image in the neighbor set). The basic intuition is that if different persons label similar images using the same tags, these tags are likely to reflect objective aspects of the visual content. In the second step, for each tag t of d, its tag relevance is the probability of t being used among the K nearest neighbors minus the probability of t being used among the image collection. Note that the association between visual features and visual similarities is a challenging problem. In a recent work, Li *et al.* compared tag relevance learned by considering visual similarities defined by multiple types of visual features and concluded that a uniform combination of neighborhood images based on multiple visual features yield comparable or better results than other combination methods [9].

In [18], three tag relevance formulations are evaluated, namely, *unit*, *tag-position*, and *neighbor-voting* relevance. The experimental results demonstrate that neighbor-voting relevance significantly outperforms other formulations for single-tag queries. It is also interesting to observe that for multi-tag queries, the choice of tag relevance does not significantly affect the TagIR accuracy probably due to the fact that a multi-tag query usually expresses a very specific information need. Hence, in this paper we evaluate the tag relevance voting schemes using single-tag queries.

Neighbor-voting based tag relevance has also been used in other related applications. Liu *et al.* re-ranked the tags of a tagged image such that the most relevant tags appear in top positions [11]. The authors used neighbor-voting as the first step and then applied random-walk to further refine the learned tag relevance by considering pair-wise similarity between tags. 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 [10, 20]. In this work, we evaluate the dimensions in tag relevance voting through comparing different voting schemes. Such an evaluation would benefit not only TagIR but also other aforementioned techniques.

K-Nearest Neighbor Classifier. Most germane to this work are studies on k-Nearest Neighbor (*kNN*) classifiers. *kNN* classifies an unseen instance based on the category labels of its nearest neighbors. In other words, each of its k nearest neighbors serves as an evidence that supports the likelihood of the instance belonging to certain category. Due to its simplicity and effectiveness, *kNN* has been widely adopted in various classification tasks and also gained research interests from multiple aspects such as the impact and choice of k, the choice of similarity function [19]. In our setting, the similarity function is defined based on the visual features used to describe the images [9].

More related to this work are the studies on the voting schemes (unweighed or weighted voting) in *kNN* with predetermined k. In weighted voting, the similarity or distance between the unseen instance and its neighbors are often used to determine the voting power a neighbor has. Clearly, in tag relevance learning, this maps to the similarity between the seed image and its neighbors. Another interesting dimension in *kNN* study is the certainty of the labeled data. Given a labeled instance, a label often serves as a binary indicator to specify whether or not this instance belongs to that class [5, 6]. However, in reality, the certainty of the label itself may not be uniform across different labels, *e.g.,* in the diagnostic domain. In [6], a fuzzy *kNN* is proposed to address the uncertainty of one instance may belong to multiple classes having different strengths. Unfortunately, in the context of social tagging, the tags are assigned by users freely with different motivations for tagging and different interpretations of the relevance between a tag and an image. Consequently, the "labels" in social tagging setting are much more uncertain compared to those in traditional classification problems.

Table 1: Table of notations.

$\mathcal{D}$	The image collection
$\mathcal{T}$	The set of all tags in $\mathcal{D}$
d	an image or the seed image whose tag relevance is learned
d'	one of the neighbor voting images
V_d	set of visual features of d
T_d	set of tags of d
t	a tag in T_d
t'	a voting tag in the neighbor image d'
$\mathbf{r}$	the tag relevance vector of an image
$\mathbf{r}_i$	i-th element of $\mathbf{r}$, the tag relevance of the i-th tag $t_i \in T_d$
$\mathbf{v}_i$	i-th element of $\mathbf{v}$, the neighbor vote of tag $t_i \in T_d$
$\mathbf{P}$	the random walk's transition probability matrix

Clearly, social image tag relevance shares much similarity with the *kNN* classification problem. On the other hand, the key difference between them is that tag relevance learning estimates the effectiveness of a user-assigned tag in describing an image, which is a continuous value (*e.g.,* normally in the range of [0,1] after normalization); however, *kNN* returns a crisp decision of a category label where some decision theory can be applied and the voting from some neighbors can be rejected [5]. Additionally, in a classification setting, the categories are predefined by domain experts and are distinctive from each other, especially in flat classification problems where the categories do not form a topic hierarchy. In contrast, in social tagging it is common to use multiple tags to describe the same image (*e.g.,* red, rose, flower for the seed image shown in Figure 1). In fact, different concepts may emerge from the tag co-occurrences [17, 18]. This provides us the opportunity to exploit additional dimensions to improve the baseline scheme for more accurate tag relevance learning as highlighted in Section 1.

3. NEIGHBOR VOTING SCHEMES

In this section, we first give an overview of the neighbor voting for tag relevance. Next, we enumerate a subset of alternative formulations for different content and context-based voting schemes. We begin by introducing some notations (see Table 1) to facilitate our discussions.

Given the image collection $\mathcal{D}$ and the set $\mathcal{T}$ of all tags in $\mathcal{D}$, an image $d \in \mathcal{D}$ is a 2-tuple $\langle V_d, T_d \rangle$, where V_d is a collection of low-level features/descriptors derived from the visual content of the image and $T_d \subseteq \mathcal{T}$ is a collection tags assigned by users. For each tag $t_i \in T_d$, the tag relevance $\mathbf{r}_i \in [0, 1]$ measures how accurately t_i objectively describes the visual content of d. The tag relevance vector $\mathbf{r}$ of d is the list of tag relevances of all d's tags, where $\mathbf{r}_i$ is the tag relevance of the i-th tag in T_d.

3.1 Tag Influence-Unaware Neighbor Voting

We first consider neighbor voting scheme without the tag influence dimension. We assume that each neighbor image d' is allowed certain amount of voting power for the relevance of a tag t for seed image d. Equation 1 computes the voting power based on the similarity between the seed image d and neighbor image d', the tag matching model between the tag t and all tags of d' (denoted by $T_{d'}$), and the effectiveness of $T_{d'}$ in describing d' (or tag relevances of d'). Observe that this generic voting scheme considers the *image similarity*, *tag matching*, and *tag relevance* dimensions.

$$vote(t, d, d') = sim(V_d, V_{d'}) \times match(t, T_{d'}) \times relevance(T_{d'}, d') \quad (1)$$

$$\mathbf{v}_i = \frac{\sum_{d' \in N(d)} vote(t_i, d, d')}{|N(d)|} - \frac{\sum_{d' \in \mathcal{D}} vote(t_i, d, d')}{|\mathcal{D}|} \quad (2)$$

The vote $\mathbf{v}_i$ of tag $t_i \in T_d$ for image d is computed by Equation 2. Intuitively, it is the voting power received from the visually similar neighbor images of d ($N(d)$) offset by the voting power expected from arbitrary images in the collection. The tag relevance of t_i is the globally normalized $\mathbf{v}_i$ in the collection.

While the computation of the voting power received from the neighbor images (the first component in Equation 2) is straightforward, the computation for the collection's *expected voting power* (the second component in Equation 2) is expensive, requiring the vote from *all* images in $\mathcal{D}$. One approach to simplify the computation is to assume that the three functions $sim(V_d, V_{d'})$, $match(t, T_{d'})$, and $relevance(T_{d'}, d')$ are independent, and to estimate them independently. However, such assumption contradicts the *neighbor-voting* assumption (visually similar images are annotated with similar tags). That is, $sim(V_d, V_{d'})$ and $match(t, T_{d'})$ are not independent. In our experiments, we therefore estimate the *expected voting power* from the collection through sampling. In each computation, 2600 images (about 1% of the entire collection) are randomly sampled from the collection for estimation.

In the following, we discuss the alternative formulations for each dimension with reference to the baseline voting scheme.

Image Similarity. In the baseline voting scheme, the voting power allocated to a neighbor image is independent of the visual similarity between the neighbor and the seed images. That is, $sim(V_d, V_{d'}) = 1$ for every neighbor image. Similar to most *kNN* rules, we propose a *weighted voting* scheme to allocate more voting power to those neighbor images that are more visually similar to the seed image. The function $sim(V_d, V_{d'})$ is then the visual similarity between images d and d' used to finding the nearest neighbors. For ease of comparison, we assume that $sim(V_d, V_{d'})$ is normalized to [0, 1].

Tag Matching. In the baseline voting scheme, a binary tag matching function is adopted. If a neighbor image d' contains tag t (*i.e.,* $t \in T_{d'}$) then its voting power to t is 1 otherwise 0. Since tags are incomplete, we propose to use *tag association measures* to allocate some amount of voting power to a neighbor image even if it does not contain tag t but contains some tags that are highly associated (or co-occurred) with t. The intuition behind this strategy is that users may use highly co-occurring tags instead of t to tag the neighbor images (*e.g.,* flower instead of rose). Note that it has been recently empirically demonstrated that tag association matching model significantly improves TagIR accuracy [18].

To allocate voting power to a neighbor image d' not containing tag t, we propose two formulations referred to as *Maximum* and *Power Mean* associations, respectively, as shown in Equation 3.

$$match(t, T_{d'}) = \begin{cases} \max_{t' \in T_{d'}} assoc(t, t') & \text{Maximum} \\ \left(\frac{1}{|T_{d'}|} \sum_{t' \in T_{d'}} assoc(t, t')^p \right)^{1/p} & \text{Power mean.} \end{cases} \quad (3)$$

The *Maximum* formulation allocates voting power to d' based on its most associative tag to t in T_d. On the other hand, the *Power Mean* formulation considers associations from all tags in d'. A large p signifies the voting power is mostly contributed by the most associative tags in $T_{d'}$ with t, while a small p indicates the contribution of all tags in $T_{d'}$ are relatively equal. Observe that the power mean association is reduced to maximum association when p approaches infinity. In our experiments, we set $p = 20$ after empirical tuning. We use Jaccard coefficient for the association measure between two tags $assoc(t, t')$ due to its relatively good performance in various social tagging tasks [16, 18].

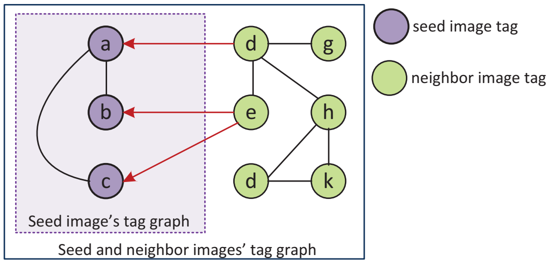


Figure 2: An example tag graph (edges without arrows are bi-directional).

Refined Tag Relevance. In the baseline voting, for a given neighbor image d', all its tags are assumed to have the same effectiveness in describing the image's visual content and given the same voting power, *i.e.,* $relevance(t', d') = 1, \forall t' \in T_{d'}$. However, tags in a neighbor image $t' \in T_{d'}$ also have different effectiveness in describing d'. We therefore propose to first learn tag relevances for all image $d \in \mathcal{D}$. Then, the learned $relevance(t', d')$ (normalized in [0, 1]) is used to re-learn tag relevance for a given seed image as in Equation 1. However, we note that noise could be easily introduced if we consider tag associative matching and tag relevance together. For instance, a tag $t' \in T_{d'}$ may be marginally relevant to the image d' but largely associated with the voting tag t. We therefore consider this refined tag relevance dimension only in the context of exact tag match.

3.2 Tag Influence-Aware Neighbor Voting

As the visual content of an image is often annotated with multiple associative tags (*e.g.,* red, rose, and flower), it is interesting to investigate whether the relevance of a group of correlated tags further boost the accuracy of tag relevance computation as these tags jointly describe the visual concept of the image. We propose two variants of the *tag influence model* using random walk by considering either the association between the tags in the seed image or the association between tags in both the seed image and the neighbor images. The tag relevance is then the result of influence voting on top of the relevance computed by neighbor voting as shown in Equation 4, where the i-th element of vector $\mathbf{v}$ is the result of neighbor voting on tag $t_i \in T_d$ (see Equation 2).

$$\mathbf{r} = influenceVote(\mathbf{v}, d) \quad (4)$$

Influence of Tags in the Seed Image. In this approach, we represent the tags associated with the seed image d as an undirected graph G where nodes are tags T_d in d. Two tags are connected with an edge with weight $assoc(t_1, t_2)$ when $assoc(t_1, t_2) > 0$. The graph in Figure 2 encompassed by dotted rectangular box is an example of such tag graph.

The *transition probability matrix* of G is denoted as $\mathbf{P}$ whose element p_{ji} indicates the probability of the transition from node i to node j. Note that $\forall i, j, p_{ji} = assoc(t_i, t_j) / \sum_j assoc(t_i, t_j)$. Let $\mathbf{v}$ be the vote vector computed from neighbor voting (Equation 4).

Let $\mathbf{r}^{(k)}$ be a vector whose i-element $\mathbf{r}_i^{(k)}$ indicates the relevance vote of tag t_i at step k. The random walk process is then formulated as follow:

$$\mathbf{r}^{(k)} = \alpha\mathbf{P}\mathbf{r}^{(k-1)} + (1-\alpha)\mathbf{v} \tag{5}$$

where $\alpha \in [0, 1)$ is a parameter indicating the weight of tag relations computed by random walk within the final computed tag relevance (a discussion on α is given at the end of this section). Mathematically, we should define the initial relevance $\mathbf{r}^{(0)}$ as well. However, as we shall prove later, in a random walk, the initial state $\mathbf{r}^{(0)}$ does not affect the stationary state. We can assign $\mathbf{r}^{(0)} = \mathbf{v}$ or $\mathbf{r}^{(0)} = (1, 1, \ldots, 1)$. From Equation 5, the relevance score at step k is given below. The reader may refer to Appendix for the proof of invertibility of $(\mathbf{I} - \alpha\mathbf{P})$.

$$\begin{aligned}\mathbf{r}^{(k)} &= (\alpha\mathbf{P})^k\mathbf{r}^{(0)} + (1-\alpha)\Big(\sum_{i=1}^{k}(\alpha\mathbf{P})^{i-1}\Big)\mathbf{v}\\ &= (\alpha\mathbf{P})^k\mathbf{r}^{(0)} + (1-\alpha)\Big(\mathbf{I}-\alpha\mathbf{P}\Big)^{-1}\Big(\mathbf{I}-(\alpha\mathbf{P})^k\Big)\mathbf{v}\end{aligned}$$

Since $\mathbf{P}$ is a non-negative column-normalized matrix, $\forall k, \mathbf{P}^k$ is also a non-negative column-normalized matrix. Subsequently, all elements in $\mathbf{P}^k$ are in $[0, 1]$ since they are probabilities. Thus, $\lim_{k\to\infty}(\alpha\mathbf{P})^k = \mathbf{0}$. Hence,

$$\mathbf{r} = \lim_{k\to\infty}\mathbf{r}^{(k)} = (1-\alpha)\Big(\mathbf{I}-\alpha\mathbf{P}\Big)^{-1}\mathbf{v}$$

Influence of Tags in Neighborhood. In this approach, instead of only considering the tags in the seed image d, all tags in both d and $N(d)$ are considered. However, since our main purpose is to compute the relevance of tags in T_d, we prevent the tag relevance of tags in T_d to "influence" the tag relevance of tags not in T_d. Consequently, the tag graph is now a directed graph $G(N_G, E_G)$. The node set N_G contains of two partitions, denoted by N_S and N_T, where N_S contains all tags in T_d and N_T contains the remaining nodes. There is an edge from t_1 to t_2 with weight of $assoc(t_1, t_2)$ if (i) $assoc(t_1, t_2) > 0$ and (ii) $t_1 \notin N_S$ or $t_2 \notin N_T$. Figure 2 illustrates the interaction between the tag graphs of seed and neighbor images. Note that the red edges indicate influences from N_T to N_S. Mathematically, the transition probability matrix $\mathbf{P}$ is described as follows:

$$\mathbf{P} = \begin{bmatrix}\mathbf{P}_S & \mathbf{P}_{ST}\\ \mathbf{0} & \mathbf{P}_T\end{bmatrix}$$

where $\mathbf{P}_S$ and $\mathbf{P}_T$ are square matrices indicating the transition probability within N_S and N_T, respectively, while $\mathbf{P}_{ST}$ stores the transition probability from N_T to N_S. Note that $\mathbf{P}_S$ is identical to the transition probability matrix in the aforementioned tag graph of seed image. Further, we follow the same random walk process as used in the above model. In particular, the random walk follows Equation 5 leading to the same convergence of $(1-\alpha)\Big(\mathbf{I}-\alpha\mathbf{P}\Big)^{-1}\mathbf{v}$.

Remark on α. Note that $0 \leq \alpha < 1$. When $\alpha = 0$, $\mathbf{r}^{(k)} = \mathbf{v}\forall k$ and tag relation is not considered. There are two reasons for setting $\alpha < 1$. First, $\alpha = 1$ means neighbor voting is not considered at all. Note that, if converged, the convergence state of random walk is dependent on only the transition matrix but not the initial state. Second, when $\alpha = 1$, the random walk used on neighborhood tag graph may not converge. According to Perron-Frobenius Theorem [14], a random walk converges when its transition matrix is irreducible and aperiodic. However, here, $\mathbf{P}$ is reducible. Hence, in our experiments we set $\alpha = 0.5$.

Table 2: Voting notations for the four dimensions.

Notation	Semantics
S_u	Unit image similarity (baseline)
S_w	Weighted image similarity
M_e	Exact tag matching (baseline)
M_x	Tag matching by maximum association
M_p	Tag matching by power mean association
I_o	Without tag influence (baseline)
I_s	Influence by tags in seed image
I_n	Influence by tags in seed and neighbor images
R_u	Unit tag relevance (baseline)
R_r	Using learned tag relevance to re-learn relevance

4. EXPERIMENTAL SETUP

In this section, we report the experimental settings for evaluating the proposed voting schemes for tag relevance learning. The experimental setting largely follows the settings in [18] and conducted on the NUS-WIDE dataset [3].

4.1 Dataset

The NUS-WIDE dataset contains 269,648 images from Flickr and 81 tags are manually labeled with ground-truth matching images. All tags provided in the dataset are used in our experiments without filtering.

Image Similarity. The NUS-WIDE dataset provides six types of low-level features to describe the visual contents of images, including global features such as color, edge, texture, and local feature known as bag of visual-words. Two sets of nearest neighbors are obtained using global and local features, respectively. For global features, 64-D color histogram, 73-D edge direction histogram, and 128-D wavelet texture features are used where the three types of features for each image were aggregated into a 265-D vector after unit-length normalization on each type of features. The nearest neighbors are determined using Euclidian distance. For local features, the 500-D bag of visual-words are used. The nearest neighbors are determined by cosine similarity with $tf \times idf$ word weighting scheme. Following [8], the neighbor images satisfy unique-user requirement to avoid the possible bias introduced by the same user annotating too many similar images for a seed image.

Number of Neighbors. To find the k nearest neighbors of a given image, we ensure that half of the neighbors are based on global features and the other half are based on local features. We set $k \in \{100, 200, 400, 1000\}$. We observed that adding more neighbors only lead to very minor improvement in the results. Since number of neighbors is not crucial for our study, we report our experimental findings with $k = 400$ (*i.e.,* 200 nearest neighbors are obtained based on global features and another 200 neighbors are based on local features). We first compute the tag relevance using the 200 neighbors by global feature and the tag relevance using the 200 neighbors by local feature, respectively. This is because the image similarities defined by global feature and local feature are not directly comparable. Next, we average the learned tag relevance.

4.2 Naming of Voting Schemes

A voting scheme is a combination of alternative formulations selected under the four proposed dimensions. We use the notations listed Table 2 to uniquely identify each formulation. For instance, $S_uM_eI_oR_u$ refers to the baseline voting scheme, with unit (or unweighed) image similarity, exact tag matching, without tag influence, and unit tag relevance for neighbor images. Since there are two choices for image similarity, three choices for tag matching, and three choices for tag influence, there will be 18 different voting schemes that need to be evaluated. For tag relevance, we evaluate

Table 3: Voting schemes ranked by $P@100$ and MAP. The "+" and "-" indicate the results are significantly better or worse than the baseline statistically. The baseline is underlined.

Rank	Scheme	$P@100$	Scheme	MAP
1	$S_wM_eI_oR_u$	0.7515+	$S_wM_eI_oR_u$	0.3660
2	$S_wM_eI_sR_u$	0.7414	$S_uM_pI_oR_u$	0.3646+
3	$S_wM_pI_oR_u$	0.7390	$S_uM_pI_sR_u$	0.3634
4	$S_wM_xI_oR_u$	0.7388	$S_wM_eI_sR_u$	0.3633
5	$S_wM_eI_nR_u$	0.7364	$\underline{S_uM_eI_oR_u}$	0.3630
6	$S_wM_pI_sR_u$	0.7300	$S_uM_eI_nR_u$	0.3621-
7	$S_uM_pI_oR_u$	0.7283+	$S_uM_eI_sR_u$	0.3619-
8	$S_wM_xI_sR_u$	0.7254	$S_uM_pI_nR_u$	0.3617
9	$S_uM_pI_sR_u$	0.7243	$S_wM_pI_oR_u$	0.3615
10	$\underline{S_uM_eI_oR_u}$	0.7231	$S_uM_xI_oR_u$	0.3611-
11	$S_uM_pI_nR_u$	0.7227	$S_wM_xI_oR_u$	0.3610
12	$S_wM_pI_nR_u$	0.7222	$S_wM_eI_nR_u$	0.3602
13	$S_wM_eI_oR_r$	0.7216	$S_uM_xI_sR_u$	0.3600-
14	$S_uM_eI_nR_u$	0.7211	$S_uM_xI_nR_u$	0.3597-
15	$S_wM_xI_nR_u$	0.7195	$S_wM_eI_oR_r$	0.3593
16	$S_uM_eI_sR_u$	0.7186-	$S_wM_pI_sR_u$	0.3588
17	$S_uM_xI_oR_u$	0.7179-	$S_wM_xI_sR_u$	0.3582
18	$S_uM_xI_nR_u$	0.7152-	$S_uM_eI_oR_r$	0.3581-
19	$S_uM_xI_sR_u$	0.7122-	$S_wM_pI_nR_u$	0.3565
20	$S_uM_eI_oR_r$	0.7064-	$S_wM_xI_nR_u$	0.3558-

the two voting schemes with exact tag matching and without tag influence because of the reason discussed earlier. In summary, 20 voting schemes (including the baseline) are evaluated by our study.

4.3 TagIR Task and Evaluation Metrics

Since tag relevance is less significant for multi-tag queries [18], we evaluate the TagIR accuracy using the 81 single-tag queries. The manually labeled images to those 81 tags serve as the ground-truth in our evaluation. Note that in [18] the best performing methods use neighbor-voting tag relevance (which will be evaluated in this work) and, either (i) associative tag matching for matching an image's tag to a query tag or (ii) query expansion. For the case of query expansion, the image retrieval system basically answers a multi-tag query expanded from a single-tag query and the results depend on the expansion techniques used. In this evaluation, we therefore consider associative tag matching where for a given single-tag query (*e.g.,* rose), the matching score of an image is computed based on its tag matching the query tag (*i.e.,* rose) and its tags that are highly associated with the query tag (*e.g.,* flower). The results of $Q_SR_VD_FL_SM_J$[1] defined in [18] is reported in this work. It is the fifth best performing method without query expansion. We selected this method because similar to our voting schemes it uses Jaccard coefficient for tag associative measure. Note that, our baseline result is slightly different from the one reported in [18] because of two reasons: (i) we apply the unique-user constraint in neighbor voting, and (ii) we apply global min-max normalization for the tag relevance values learned in the entire data collection while [18] applied min-max normalization to the learned tag relevances within every image.

We report the performance of the voting schemes on this TagIR task using two measures, *Precision@K* (*P@K*) and *Mean Average Precision* (MAP). Note that $P@K$ may better evaluate a user's perception about a TagIR system as in reality a keyword query may match a large number of images and a user is unlikely to go through all returned images. For example, there are more than 74K images matching query sky in the NUS-WIDE dataset; queries like clouds, person, and water, each matches more than 30K images.

Table 4: Impact of the image similarity dimension.

Scheme	$P@100$	MAP
$M_eI_oR_u$	$S_u \ll S_w$	$S_u \approx S_w$
$M_xI_oR_u$	$S_u \approx S_w$	$S_u \approx S_w$
$M_pI_oR_u$	$S_u \approx S_w$	$S_u \approx S_w$
$M_eI_sR_u$	$S_u \ll S_w$	$S_u \approx S_w$
$M_xI_sR_u$	$S_u \approx S_w$	$S_u \approx S_w$
$M_pI_sR_u$	$S_u \approx S_w$	$S_u \approx S_w$
$M_eI_nR_u$	$S_u \approx S_w$	$S_u \approx S_w$
$M_xI_nR_u$	$S_u \approx S_w$	$S_u \approx S_w$
$M_pI_nR_u$	$S_u \approx S_w$	$S_u \approx S_w$
$M_eI_oR_r$	$S_u \approx S_w$	$S_u \approx S_w$

Precision@K is the ratio of the relevant images of the top-K retrieved images for a sample query. We measured different $K \in \{25, 50, 100, 200, 400\}$ for the 81 queries and observed that $P@K$ values for different values of K follow very similar trend. Due to space constraints, we choose to report $P@100$ only. For each method, the reported $P@100$ is the macro-average of $P@100$ values of all evaluated queries.

Mean Average Precision. For each query, Average Precision measures the average precision values obtained when each relevant image is retrieved [12]. Mean Average Precision (MAP) is the mean of the Average Precisions for all sample queries.

5. RESULTS AND DISCUSSIONS

In this section, we first present an overview of the empirical results related to the 20 voting schemes on $P@100$ and MAP. Next, we give a detailed comparison of our results over the four dimensions. Finally we summarize our empirical findings.

5.1 Overview

Table 3 ranks the 20 voting schemes by $P@100$ and MAP respectively. Observe that, by $P@100$ the baseline method is ranked at the 10th position. Among the 9 methods outperform the baseline, 7 of them use weighted image similarity. Notably, the only differentiating factor of the best performing voting scheme $S_wM_eI_oR_u$ compared to the baseline $S_uM_eI_oR_u$ is S_w over S_u (The performance difference is significant by paired t-test). Interestingly, tag matching model seems to play a less significant role as there is no clear pattern among the 9 methods outperformed the baseline. Furthermore, the two voting schemes related to tag relevance perform poorer than the baseline. In particular the method $S_uM_eI_oR_r$, differs only in tag relevance dimension from the baseline, performs the worst among all methods and is significantly worse than the baseline.

In summary, weighted voting based on image similarity is a better choice than unweighed voting; refinement of tag relevances of neighbor images actually hurts the performance; incorporating tag matching model or tag influence voting has insignificant impact on the results.

Now consider the ranking based on MAP. The baseline method now occupies the 5th position. The method $S_wM_eI_oR_u$ remains the best performing method.

5.2 Alternative Choices vs Baseline

It is evident from the aforementioned results that only few voting schemes significantly outperform the baseline. In this section, we present a detailed analysis on the impact of using alternative formulations under each dimension against the baseline.

Image Similarity. Table 4 reports the comparison between S_w and S_u used by baseline. For any pair of voting schemes that only dif-

[1] This method uses single query without query expansion (Q_S), neighbor-voting based tag relevance (R_V), inverse document frequency for tag weighting (D_F), Lucene's default square root length normalization (L_S), and Jaccard coefficient for tag associative matching (M_J) between the image tag and query tag.

Table 5: Impact of the tag matching dimension.

Scheme	$P@100$		MAP	
$S_uI_oR_u$	$M_e \gg M_x$	$M_e \ll M_p$	$M_e \gg M_x$	$M_e \ll M_p$
$S_wI_oR_u$	$M_e \gg M_x$	$M_e \gg M_p$	$M_e \gg M_x$	$M_e \approx M_p$
$S_uI_sR_u$	$M_e \gg M_x$	$M_e \ll M_p$	$M_e \gg M_x$	$M_e \ll M_p$
$S_wI_sR_u$	$M_e \gg M_x$	$M_e \gg M_p$	$M_e \gg M_x$	$M_e \gg M_p$
$S_uI_nR_u$	$M_e \gg M_x$	$M_e \ll M_p$	$M_e \gg M_x$	$M_e \gg M_p$
$S_wI_nR_u$	$M_e \gg M_x$	$M_e \approx M_p$	$M_e \gg M_x$	$M_e \gg M_p$

Table 6: Impact of the tag influence dimension.

Scheme	$P@100$		MAP	
$S_uM_eR_u$	$I_o \gg I_s$	$I_o \approx I_n$	$I_o \gg I_s$	$I_o \gg I_n$
$S_wM_eR_u$	$I_o \gg I_s$	$I_o \gg I_n$	$I_o \gg I_s$	$I_o \gg I_n$
$S_uM_xR_u$	$I_o \gg I_s$	$I_o \approx I_n$	$I_o \gg I_s$	$I_o \gg I_n$
$S_wM_xR_u$	$I_o \gg I_s$	$I_o \gg I_n$	$I_o \gg I_s$	$I_o \gg I_n$
$S_uM_pR_u$	$I_o \gg I_s$	$I_o \gg I_n$	$I_o \gg I_s$	$I_o \gg I_n$
$S_wM_pR_u$	$I_o \gg I_s$	$I_o \gg I_n$	$I_o \gg I_s$	$I_o \gg I_n$

fer on image similarity, paired t-test is conducted based on the results on $P@100$ and MAP respectively. For example the row representing the scheme $M_eI_oR_u$ reports the paired t-test results between $S_uM_eI_oR_u$ and $S_wM_eI_oR_u$ where the symbols $\gg$, $\ll$, and $\approx$ denote statistically significant better, significantly worse, and comparable, respectively, based on the results of the 81 queries. The weighted option S_w based on $P@100$ performs significantly better than its unweighed counterpart S_u for two voting schemes and is comparably for other schemes. However, the two formulations are comparable when MAP is used.

Tag Matching. Table 5 reports the comparison of the two associative tag matching formulations, M_x and M_p, against the baseline M_e. Based on $P@100$, M_e clearly outperforms M_x. Between M_e and M_p, on clear pattern can be observed. In other words, there is no evidence that the M_p option significantly outperform the baseline M_e even with careful parameter tuning in M_p. In other words, associative tag matching performs poorly in comparison to exact tag matching. The same observation is made for the comparison based on MAP.

Tag Influence and Refined Tag Relevance. Tables 6 and 7 report the results on tag influence and refined tag relevance dimensions, respectively. Observe that the baseline formulations significantly outperform these alternative formulations. In other words, tag influence or tag relevance in neighbor voting hurts accuracy of the learned tag relevance in TagIR.

5.3 Content is Still King

Based on the above results and discussion, we can conclude that the baseline voting scheme, despite its simplicity, remains one of the most competitive scheme. Interestingly, despite the potential of tags in improving social image search, *significant improvement is achieved only by the content-based dimension* (visual similarity between the seed image and neighbor images in the voting) instead of the three context-based dimensions. Why this is so? In the following, we give our interpretations to this phenomenon.

We observed that the weighted image similarity formulation, compared to the unweighed version, better reflects the observation that visually similar images are annotated with similar tags. Hence if a neighbor image is less visually similar to the seed image, the neighbor image should be allocated with less voting power as the tags assigned to this neighbor image may not necessarily be useful for learning tag relevance for the seed image.

Table 7: Impact of refining tag relevance.

Scheme	$P@100$	MAP
$S_uM_eI_o$	$R_u \gg R_r$	$R_u \gg R_r$
$S_wM_eI_o$	$R_u \gg R_r$	$R_u \gg R_r$

On the other hand, the relatively poor contributions of the three context-based dimensions (tag matching, tag influence, tag relevance) stem from the nature of tagging behavior of users. For instance, the associative tag matching dimension is proposed based on the assumption that users may not tag visually similar image using exactly the same tag t. Instead, they may use another tag t' which is highly associated with t. Hence, in this formulation t will receive some voting power from a neighbor image having t' but not t. An aftermath of this formulation is that it makes a tag t, which is irrelevant to the seed image d, to gain more voting from the neighbor images. On the other hand, due to the unique user constraint for nearest neighbors selection, each neighbor image are from a different user. Given the large number of neighbors considered in voting (*e.g.,* 400 in our experiments), it is highly unlikely that this large number of users would miss out a very relevant tag.

Now consider the tag influence dimension. The alternative formulations of this dimension attempt to further boost tag relevance by considering tag co-occurrences. Note that the sampling of nearest neighbors by visual similarity is independent of user tagging behavior. In reality, if tags t_1 and t_2 co-occur often in the whole data collection, they are likely to co-occur often in the neighbor images of any particular seed image. In other words, if t_1 receives more voting, then t_2 would likely to receive more voting as well. Consequently, the ability to further boost t_1 or t_2's voting because of their high co-occurrence diminishes and the tag influence dimension becomes less effective in TagIR.

Lastly, consider the impact of using learned tag relevance in the first iteration to refine the learning in the next iteration. We evaluated two methods with exact tag matching without tag influence, and to our surprise we observed that the results are both negative. A closer look at the results reveals that it is possible to get slight improvement for some queries which also achieve relatively good results with the baseline voting scheme ($P@100$ or MAP). However, for queries that perform poorly with the baseline voting scheme, there is significant drop in performance in the presence of this dimension.

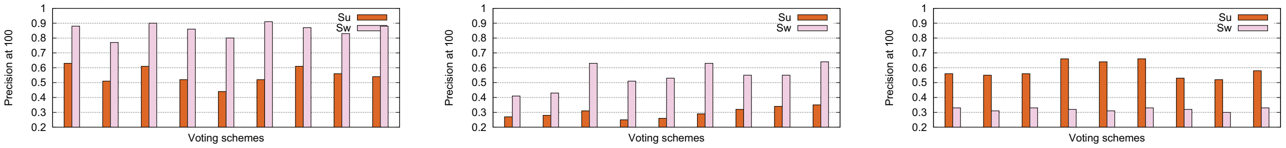
5.4 Case Study

We observe that for some queries (*e.g.,* animal, horses, plants, tiger, plane, toy, sky, clouds, coral, dog) the voting schemes have minor impact on the retrieval accuracy. On the other hand, they have significant impact on the retrieval accuracy of some other queries (*e.g.,* military, cat, valley, train, fire, nighttime, surf, rainbow, cars, garden, house).

Among the 81 queries evaluated, Figure 3 shows the $P@100$ results for three single-tag queries (military, cat, and valley) having largest variances among the 20 voting schemes. Clearly, the choice of image similarity formulation in the voting schemes play pivot role for TagIR performance. Specifically, the weighted voting benefits queries like military and valley but adversely affect the performance of cat. However, no clear pattern related to the "types" of queries that would benefit from weighted or unweighed voting can be observed from the experimental results.

6. CONCLUSION AND FUTURE WORK

The availability of user-given tags as meta-data has given rise to opportunities to build novel and superior tag-based techniques to significantly enhance our ability to understand social images and to retrieve them effectively and efficiently. One of the key challenge in this context is determining how accurately a tag objectively re-



(a) Query military (b) Query valley (c) Query cat

Figure 3: $P@100$ for single-tag queries: military, valley, and cat.

flect the visual content of the image, known as tag relevance. In this paper, we explore the effectiveness of different variants of the neighbor voting technique for tag relevance learning. Specifically, we investigate a content-based dimension (image similarity) and three context-based ones (tag matching, refined tag relevance, and tag influence) that might affect the accuracy of tag relevance learning. We exhaustively explored 20 neighbor voting schemes based on these dimensions with 81 single-tag queries on NUS-WIDE dataset. Surprisingly, our results reveal that significant improvement of accuracy in tag relevance learning for TagIR is achieved only by the content-based dimension instead of the three context-based dimensions.

There are two interesting issues that we intend to explore in the future. First, our results show that weighted voting significantly outperform unweighed voting based on image visual similarity. However, it also shows that the weighted voting benefit some queries significantly but hurt the performance of some other queries. It is interesting to explore the possibility of a *query-dependent* voting scheme that can *apriori* select weighted or unweighed voting depending on the characteristics of the query (or the tags to be voted in the neighbor voting). Second, we observe that the estimation of visual similarity of a seed image to the entire collection may affect the results. It is part of our future work to explore different mechanisms to offset the tag voting expected from the image collection. In summary, the results of this paper are an important first step in this regard.

7. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proc. SIGCHI*, pages 971–980. ACM, 2007.

[2] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:394–410, 2007.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proc. CIVR*, pages 48:1–48:9, 2009.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), 2008.

[5] T. Denceux. A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE Trans. On Systems Man And Cybernetics*, 25(5):804–813, 1995.

[6] J. M. Keller, M. R. Gray, and J. Givens. A fuzzy k nearest neighbor algorithm. *IEEE Trans. on Systems, Man, and Cybernetics*, SMC-15(4):580–585, July/August 1985.

[7] X. Li, C. G. M. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Proc. MIR*, pages 180–187. ACM, 2008.

[8] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Trans. Multimedia*, 11(7):1310–1322, 2009.

[9] X. Li, C. G. M. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proc. CIVR*, pages 10–17. ACM, 2010.

[10] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Image retagging. In *Proc. ACM Multimedia*, pages 491–500, 2010.

[11] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proc. WWW*, pages 351–360. ACM, 2009.

[12] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. HYPERTEXT*, pages 31–40. ACM, 2006.

[14] C. Meyer. *Matrix analysis and applied linear algebra*. SIAM, 2000.

[15] S. Overell, B. Sigurbjörnsson, and R. van Zwol. Classifying tags using open content resources. In *Proc. WSDM*, pages 64–73. ACM, 2009.

[16] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. WWW*, pages 327–336. ACM, 2008.

[17] A. Sun, S. S. Bhowmick, and J.-A. Chong. Social image tag recommendation by concept matching. In *Proc. ACM Multimedia*, pages 1181–1184. ACM, 2011.

[18] A. Sun, S. S. Bhowmick, K. T. Nam Nguyen, and G. Bai. Tag-based social image retrieval: An empirical evaluation. *Journal of the American Society for Information Science and Technology (JASIST)*, 62(12):2364–2381, 2011.

[19] H. Wang. Nearest neighbors by neighborhood counting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:942–953, June 2006.

[20] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proc. ACM Multimedia*, pages 461–470. ACM, 2010.

APPENDIX

Lemma 1. *For all* $\mathbf{P}$ *and* α, $\mathbf{I} - \alpha\mathbf{P}$ *is invertible.*

Proof. Note that $\mathbf{I} - \alpha\mathbf{P} = (\mathbf{I} - \alpha\mathbf{P}^T)^T$ and the transpose of an invertible matrix is invertible. Thus, we shall prove that $\mathbf{I} - \alpha\mathbf{P}^T$ is invertible. It is equivalent with $(\mathbf{I} - \alpha\mathbf{P}^T)\mathbf{x} = \mathbf{0}$ only has a trivial solution $\mathbf{x} = \mathbf{0}$.

$$\begin{aligned} (\mathbf{I} - \alpha\mathbf{P}^T)\mathbf{x} &= \mathbf{0} \\ \mathbf{x} &= \alpha\mathbf{P}^T\mathbf{x} \\ \mathbf{x}_i &= \sum_j \alpha p_{ji}\mathbf{x}_j \end{aligned}$$

Note that $0 \leq p_{ji} \leq 1 \forall i, j$ and $\sum_j p_{ji} = 1 \forall j$. Let $m = arg \min\{\mathbf{x}_i\}$, $\mathbf{x}_m = \sum_j \alpha p_{jm}\mathbf{x}_j \geq \sum_j \alpha p_{jm}\mathbf{x}_m = \alpha\mathbf{x}_m$. Thus, $(1 - \alpha)\mathbf{x}_m \geq 0$ and $\mathbf{x}_m \geq 0$. Similarly, let $M = arg \max\{\mathbf{x}_i\}$, we also have $(1 - \alpha)\mathbf{x}_M \leq 0$ and $\mathbf{x}_M \leq 0$. Thus, $\mathbf{x}_i = 0 \forall i$. □