Quantifying Tag Representativeness of Visual Content of Social Images

Aixin Sun
School of Computer Engineering
Nanyang Technological University
Singapore 639798
axsun@ntu.edu.sg

Sourav S. Bhowmick
School of Computer Engineering
Nanyang Technological University
Singapore 639798
assourav@ntu.edu.sg

ABSTRACT

Social tags describe images from many aspects including the visual content observable from the images, the context and usage of images, user opinions and others. Not all tags are therefore useful for image search and are appropriate for tag recommendation with respect to visual content of images. However, the relationship between a given tag and the visual content of its tagged images are largely ignored in existing studies on tags and in tagging applications. In this paper, we bridge the two orthogonal areas of social image tagging and query performance prediction in Web search, to quantify tag representativeness of the visual content presented in the annotated images, which is also known as *tag visual-representativeness*. In simple words, tag visual-representativeness characterizes the effectiveness of a tag in describing the visual content of the set of images annotated by the tag. A tag is visually representative if its annotated images are visually similar to each other, containing a common visual concept such as an object or a scene. We propose two distance metrics, namely *cohesion* and *separation*, to quantify tag visual-representativeness from the set of images annotated by a tag and the entire image collection. Through extensive experiments on a subset of *Flickr* images, we demonstrate the characteristics of seven variants of the distance metrics derived from different low-level image representations and show that the visually representative tags can be identified with high precision. Importantly, these proposed distance measures are parameter free with linear or constant computational complexity, thus are effective for practical applications.

Categories and Subject Descriptors

H.2.4 [**Database Management**]: Systems—*Multimedia databases*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

General Terms

Algorithms, Experimentation

MM'10, October 25–29, 2010, Firenze, Italy.

Keywords

Tag representativeness, visual content, Flickr, social image

1. INTRODUCTION

The prevalence of digital photography devices (e.g., digital cameras, mobile phones) has lead to huge volume of images accessible online. As a result, there is an increasing research and commercial interests in building effective search mechanisms for superior image retrieval experience. Most commercial search engines adopt similar interface for image and document search, where users' information needs are specified in the form of textual queries. In this query framework, images are assumed to be well annotated with their presented visual concepts (e.g., a scene or an object). However, in reality high quality annotations by experts are generally unavailable. As a result, there is growing research interest in exploiting increasingly available social tags as well as the text segments surrounding images in Web pages to facilitate high quality image annotation [8, 10, 15].

There are three major approaches to annotate images. *Model-based* approach requires models to be trained for a predefined set of visual concepts using labeled examples. The trained models are then used to annotate new images according to their relevance to the concepts [10]. *Example-based* approach assumes that visually similar images are annotated by a similar set of tags. For a given image, tags are recommended among those associated with its nearest neighbors by visual content similarity [12]. *Knowledge-based* approach typically does not consider the visual content of images. Instead, it relies on the relationships (e.g., co-occurrence) among tags [16]. In model-based approach, the set of visual concepts are usually manually selected such that they are relatively easy to model (e.g., water, lake, building) [3, 8, 15]. Methods in the latter two approaches often treat tags *uniformly*. That is, they do not attempt to differentiate between tags used to describe visual content of images or other aspects. However, recent studies showed that tags are often noisy and imprecise in nature [3]. Consequently, as we shall discuss in the next section, not all tags contributed by common users describe the visual content of images.

1.1 Motivation

Tags are from uncontrolled vocabulary and users tag web resources including images from many different perspectives. A recent study on the types of tags across different social platforms reports that *Flickr* tags usually describe images with topic (or visual content), time, location, opinions, or self-reference [1]. *That is, not all tags contributed by users are representative of the visual content presented in the images.* For example, consider a photo of the Forbidden City uploaded by Sara which she took using her

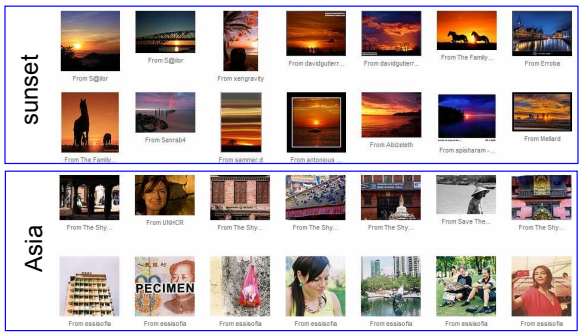


Figure 1: *Flickr* tag search results for search tags sunset **and** Asia**. The images returned for visual-representative tag** sunset **are visually coherent but not the images returned for tag** Asia**.**

Canon 40D camera when she traveled to Beijing in 2009. This image may be annotated by tags such as Canon, 40D, 2009, forbidden city, travel, Beijing, and Asia. Notice that tags like 2009 and Asia do not effectively describe the visual content of the image.

Although various studies have been conducted on image tags in recent times, surprisingly, little attention has been paid to *quantitative study of tag visual-representativeness*. Intuitively, a tag is *visually representative* if it effectively describes the visual content of its annotated images (e.g., all images share a common scene, or contain a common object). A visually representative tag (such as sunset, sky, tiger) easily suggests the scene or object an image may describe even before the image is presented to a user. On the other hand, tags like 2009, Asia, and Canon often fail to suggest anything meaningful related to the visual content of the annotated image. In other words, images annotated by a visually representative tag are visually coherent by containing similar visual content. Figure 1 captures fragments of images returned by *Flickr* for search tags sunset and Asia. Notice that there exists significant difference between the visual coherence of the images returned in response to these two search tags.

Quantifying tag visual-representativeness is an important problem and would offer twofold benefits to image retrieval. First, the knowledge of tag visual-representativeness hints the search engine about users' intention of a tag query. Users may search for images *relevant* to the visual concept described by a tag (e.g., sunset, beach); or search for images *related* to a tag if the tag does not effectively describe a visual concept (e.g., Asia). More diversified image search results are therefore expected for the latter case. Second, quantifying visual-representativeness would benefit image annotation through all aforementioned annotation approaches (model-based, example-based, knowledge-based). For model-based approach, it serves as a guideline on the selection of visual concepts to be modeled as well as on the expected accuracy of the learned model. Specifically, tags that are highly representative of visual concepts (e.g., sunset, ocean) can enable learning of more accurate models. For example-based and knowledge-based approaches, tag visual-representativeness also serves as guideline on tag recommendation, particularly when images are available with content only. In the absence of any context information, tags for camera brand (e.g., Canon) or opinions (e.g., lonely) may be inappropriate to be recommended even if the nearest neighbors happen to share these tags.

1.2 Overview

Increasingly, tags are commonly used as queries to retrieve the tagged resources such as images in *Flickr*. Analogous to the setting in Web search, in this paper, we model each tag t as a query and its tagged images I_t as the matching documents. Based on this model, we take a novel approach to *compute tag visual-representativeness by exploiting query performance prediction techniques in the Web search arena*. The intuition behind this strategy is as follows. Analogous to query performance prediction, if images in I_t all share a similar visual concept v (be it a scene or an object), then I_t is visually coherent (see Figure 1 for tag sunset as an example). We can say that all users have implicitly developed consensus on the annotation of tag t to images presenting the common visual concept v; hence t is visually representative. If images in I_t are not visually similar to each other, but more like a randomly drawn sample from an image collection, then tag t is unlikely to describe any specific visual concept, hence not visually representative. An example of such tag is Asia as depicted in Figure 1.

It is worth mentioning that techniques proposed for query performance prediction cannot be *directly* adopted to quantify visual-representativeness of image tags. Tags are from textual space while images are described in visual space. In query performance prediction (see Section 2.1 for more details), queries literally appear in documents. Consequently, a relevance score between the query and each document can be computed by the adopted document retrieval model. Such a relevance function can no longer be applied to tagged image search. Moreover, tags are assigned by different users probably with different criteria for determining the degree of relevance of an image to a tag. Another hurdle is the semantic gap between the high-level visual concepts and the low-level features extracted from images [15, 23]. That is, the low-level features may not effectively represent the visual semantics of images. Such a semantic gap does not exist between textual queries and documents, as they are in the same feature space.

We propose two distance metrics, namely *cohesion* and *separation*, to quantify the visual-representativeness of a tag by measuring (a) how well the set of tagged images presents similar visual content among them, and (b) how distinct the common visual content is with respect to the entire image collection. As we shall show later, these two metrics are generic and can be plugged into different distance functions and different image feature representations for computing tag visual-representativeness. Importantly, these measures are parameter free with linear or constant computational complexity.

Our experimental study with the NUS-WIDE dataset [3] containing images from *Flickr*, demonstrates that the proposed measures can effectively identify and rank visually representative tags. In particular, separation-based measures are more effective in identifying such tags. In summary, the major contributions of this work are as follows:

- We bridge the two orthogonal areas of social image tagging and query performance prediction in Web search, to quantify visual-representativeness of image tags. Our effort also paves way to many existing solutions for query performance prediction to be further extended to address the problem of tag visual-representativeness as well as other problems in tagging.

- In Section 3, we propose to use two distance metrics, namely *cohesion* and *separation*, to measure visual-representativeness of social tags.

- In Section 4, we compare seven variants of the proposed metrics with extensive experiments on (partial-) ground truth data and popular tags in a large image collection. We demonstrate the effectiveness of our approach to quantify visually

representative tags. Specifically, we propose a *coverage* measure to reflect the effectiveness of the visually representative tags identified by a given method for image annotation.

2. RELATED WORK

2.1 Query Performance Prediction

In the preceding section, we highlighted that our proposed approach for identifying visually representative tags exploits the techniques in the area of query performance prediction in Web search. Hence, we first define query performance prediction and then compare our approach with existing work on query performance prediction.

The goal of query performance prediction in Web search is to predict the effectiveness of a query in retrieving topically coherent documents from a collection. Given a query, if the retrieved documents are topically similar to each other, then the query is effective or unambiguous; if the retrieved documents cover various different topics, then the query is less effective or ambiguous. For instance, an unambiguous query "Flickr" submitted to Google[1] leads to top-ranked pages on *Flickr website*, *Flickr API services*, *Wikipedia entry of Flickr*, *Flickr Blog* as well as *Flickr for mobile devices*. All these pages are relevant to the photo-sharing website and its services. However, an ambiguous query "2008" to Google leads to top-ranked pages covering the following topics: *Wikipedia entry of 2008 listing the major events in 2008*, *the popular movies released in 2008 from IMDb*, *Year 2008 Calendar of United States*, and *website of the Beijing 2008 Olympic Games*. Clearly the pages for query "2008" are not topically similar to each other.

Query performance prediction enables search engines to better answer poor performing queries through alternative strategies [6, 21, 26]. While our main focus is not the alternative strategies, in the following, we discuss the major approaches for query performance prediction. One of the significant direction in this area is the computation of *query clarity score* [4]. For a given query, its *clarity score* is the Kullback-Leibler (KL) divergence between the language model estimated from the top-ranked retrieved documents by the query and the language model estimated from the entire document collection. A query is topical-specific or unambiguous if the distance is large. That is, the retrieved documents contain unusually large probabilities of words specific to a topic, such as the words photo, photography, sharing, online, and community for the query "Flickr". Observe that clarity score is analogous to the distance $Dist(Q,\mathcal{D})$ studied in the topic difficulty model proposed in [2]. A topic, denoted by $(Q,R|\mathcal{D})$, is defined by a set of queries Q reflecting the information need, and a set of relevant documents R satisfying Q, where R is drawn from collection $\mathcal{D}$. In [2], five distance measures were studied and among them, $Dist(R,\mathcal{D})$ was the most effective distance in predicting query performance, followed by $Dist(Q,\mathcal{D})$. Both $Dist(R,\mathcal{D})$ and $Dist(Q,\mathcal{D})$ were computed using Jensen-Shannon divergence between the centroids of sets Q, R, and $\mathcal{D}$ respectively.

In our earlier work [18], we first introduced *tag clarity* in the context of tagging behavior study in blogs where a tag language model is estimated from the blog posts associated with the tag and the collection language model from all blog posts. As mentioned in the preceding section, the above techniques cannot be adopted directly for quantifying tag visual-representativeness as tags are from textual space while images are described in visual space. Our first attempt of quantifying tag visual-representativeness using image tag clarity was reported in [17], which however was not well evaluated. Observe that image tag clarity corresponds to one of the 7 measures proposed in this paper, i.e., clarity-based separation with slightly different modelings of the tag clarity (see Section 3.2.2). In this work, another 6 measures are proposed to quantify tag visual-representativeness and more importantly all the 7 measures are evaluated and compared through two sets of experiments.

[1] The search results are obtained during March 2010 from www.google.com

2.2 Social Images and Tags

Recent years have witnessed increasing research efforts to study images annotated with tags in social media sharing web sites like *Flickr*. Tag recommendation, tag ranking, and tag-based classification are identified as key research tasks in this context [3]. However, only few work exploit the relationship between a tag and the content of its annotated images.

Very recently, *Flickr distance* was proposed by Wu et al. [25] to model two tags' similarity based on their annotated images. For each tag, a visual language model is constructed from 1000 images annotated with the tag and the Flickr distance between the two tags is computed using the Jensen-Shannon Divergence. Our work is significantly different from this effort in two key aspects. First, our research objective is to measure the visual-representativeness of a single tag, not the relationship between tag pairs. Second, we analyze the impact of tag frequency in its language modeling. In contrast, a fixed number (i.e., 1000) of images for each tag were sampled by Wu et al. for estimating its language model.

In [24], a probabilistic framework was proposed to resolve *tag ambiguity* in *Flickr* by suggesting semantic-orthogonal tags from those tags that co-occurred with the given set of tags. Although tag ambiguity is highly related to our work, the problem targeted in [24] is image-specific and the ambiguity is defined for a set of tags. In their problem definition, an image is assumed to be annotated by a set of tags T and T is ambiguous if there exist another two tags t_a and t_b such that adding either one gives rise to very different distributions over the remaining tags. Other than the significant difference on problem definition to our work, the solution proposed in [24] was purely based on tag co-occurrence without considering the content of annotated images.

More germane to this work is the recent efforts in measuring tag relevance to image content using *neighbor voting* [13, 12]. For a given image and its annotated tags, the *relevance* between the image and each tag is estimated through kernel density estimation in [13]. In [11], for a given image, it's k nearest neighbors are obtained by computing visual similarity through low-level features. Tags that frequently appear among the nearest neighbors (with respect to the tags' prior distribution among all images) are considered relevant to the given image. The visual relevance of a tag to an image is therefore image-specific whereas in our case, the visual-representativeness of a tag is a *global* measure and is independent of any particular images. Consequently, the two approaches are not directly comparable.

Our work is also related to [15] where the main focus is to search for high-level concepts (e.g., sunset) with little semantic gaps with respect to image representation in visual space. In [15], for a given image, its confidence score is derived based on the coherence degree of its nearest neighbors in both *visual* and *textual* spaces, assuming that each image is surrounded by textual descriptions. The high-level concepts are then derived through clustering those images with high confidence scores. Similar approach was adopted in [20]. In contrast, our approach of tag visual-representativeness computation is based on the distances between the tagged images and the collection in visual space only and not in textual space. It does not involve computationally costly nearest neighbor search or

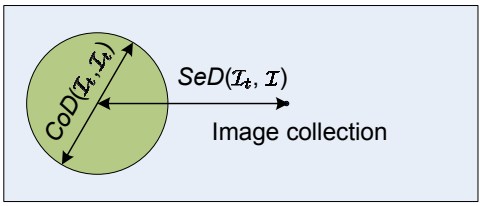


Figure 2: Illustration of cohesion and separation distances, where the circle in green denotes the set of images I_t annotated by tag t_v, and the rectangle denotes the image collection I.

clustering. Furthermore, the above approaches need to re-compute from scratch when the underlying image collection is updated. In our proposed technique, such expensive re-computation is not required as only the affected tags need to be recomputed.

3. TAG VISUAL-REPRESENTATIVENESS

In this section, we present our technique for quantifying visual-representativeness of tags in a social image collection. We begin by introducing two assumptions that we adopt for computing the tag visual-representativeness. Table 1 lists the main symbols that we use throughout the paper. We use upper case letters in calligraphic fonts for sets (e.g., I) and lower case letters for a tag or a visual concept.

Given a user-tagged image collection I, let t_v be a tag. Without loss of generality, we assume each tag t_v may describe a visual concept v by the semantic meaning of t_v. Let $I_t \subset I$ be the set of images tagged by t_v. Our main objective in this paper is to evaluate the visual-representativeness of t_v by using two distance measures as shown in Figure 2. These two measures are analogous to *cohesion* and *separation* measures in clustering evaluation [19].

- *Cohesion distance*, denoted by $CoD(I_t, I_t)$, is the distance among images in I_t. This corresponds to intra-cluster cohesion (or *compactness*) in clustering evaluation.
- *Separation distance*, denoted by $SeD(I_t, I)$, is the distance between images in I_t and I. This is similar to the inter-cluster separation (or *isolation*) measure in clustering evaluation.

In the sequel, we shall justify the reasons for considering these two measures and how they can be computed.

3.1 Assumptions

With the above notations in mind, we make the following assumptions. We shall empirically justify these assumptions in Section 4.

ASSUMPTION 1. **(Tagged images)**
If a tag t_v well describes a visual concept v, then the probability of observing v among images in I_t is larger than the probability of observing v among all images in I.

Let $\mathcal{R}_v$ be the set of images containing visual concept v. Assumption 1 states that $\frac{|I_t \cap \mathcal{R}_v|}{|I_t|} > \frac{|\mathcal{R}_v|}{|I|}$ if tag t_v well describes a visual concept. This assumption is similar to the assumption made in [12] where it is assumed that in a user-tagged dataset, the probability of correct tagging is larger than the probability of incorrect tagging. That is, it is assumed that users have done a reasonably good job in image tagging. If a user's tagging intention is to describe the visual concept(s) in a given image, then we assume that the user often selects relevant tags. Nevertheless, it is well understood that a user may tag an image from multiple aspects such as time and location of the picture other than the visual content.

Table 1: Symbols and semantics

Symbol	Semantic
I	a collection of user-tagged images
i	$i \in I$ is an image in the given collection
v	a visual concept
t_v	a tag that might describe a visual concept v
I_t	$I_t \subset I$, the set of images tagged by tag t_v
$\mathcal{R}_v$	$\mathcal{R}_v \subset I$, the set of images relevant to v

ASSUMPTION 2. **(Low-level feature distance)**
Distance derived from low-level feature representations of images reflects image visual similarity. That is, low-level feature distance among images sharing similar visual content is smaller than the distance among images not sharing similar visual content.

In content-based image retrieval (CBIR), various low-level features have been proposed for indexing images as well as measuring visual similarity between images [3, 5, 14]. These low-level features include color, texture, shapes, and others. Consequently, based on the significant research efforts in CBIR, it is reasonable to assume that the commonly-used low-level features could effectively measure visual content similarity between images.

3.2 Distance Measures

With the above assumptions in mind, we now quantify visual-representativeness of tags by revisiting the two distance measures illustrated in Figure 2. We begin by justifying the reason for choosing these two distance measures for tag visual-representativeness computation.

For a given visual concept v, let $I_{rand} \subset I$ be a randomly drawn subset from I such that $|I_{rand}| = |\mathcal{R}_v|$. Recall that all images in R_v share the common visual concept v. Hence, images in R_v are more visually similar to each other than those images in the randomly drawn subset I_{rand}. Based on Assumption 2 and distances computed using low-level features of images, $CoD(I_t, I_t) \leq CoD(I_{rand}, I_{rand})$. For the same reason, the set of images sharing the same visual concept is expected to be more distinct from the entire collection than a randomly sampled subset. That is, $SeD(\mathcal{R}_v, I) \geq SeD(I_{rand}, I)$.

In reality, $\mathcal{R}_v$ is usually unavailable due to lack of high-quality annotations by experts. However, based on Assumption 1, I_t can be a reasonably good approximation of $\mathcal{R}_v$ if the tag t_v well describes the visual concept v. In other words, if a tag t_v describes visual concept v well, then $CoD(I_t, I_t) \leq CoD(I_{rand}, I_{rand})$ and $SeD(I_t, I) \geq SeD(I_{rand}, I)$ both hold. In contrast, if a tag t does not describe a specific visual concept, then I_t becomes an approximation of I_{rand} with respect to the visual content of images. For example, tags like 2009 and Asia are very unlikely to describe any specific visual content of images. In this case, $CoD(I_t, I_t) \approx CoD(I_{rand}, I_{rand})$ and $SeD(I_t, I) \approx SeD(I_{rand}, I)$. We therefore utilize the two distances to quantify the visual-representativeness of social tags.

3.2.1 Cohesion Distance Computation

We adopt the centroid (or prototype)-based and link-based cohesion measures both commonly used in clustering evaluation [19] to compute cohesion distance among images.

Centroid-based cohesion. Let $Cent(I_t)$ be the centroid of I_t. Let *dist* be a distance function for vector representations of images or

centroids. Then the expected distance between i and $Cent(I_t)$, denoted by $\Phi_{cent}(I_t, I_t)$, can be computed as follows.

$$\Phi_{cent}(I_t, I_t) = \frac{1}{|I_t|}\sum_{i \in I_t} dist(i, Cent(I_t)) \quad (1)$$

In the above equation, depending on the types of low-level features, different distance functions may be applied [22]. In our study, we use cosine distance[2] and Euclidian distance for local (i.e., bag of visual-words) and global feature representations (i.e., color, edge, texture) of images, respectively. Note that the time complexity of the centroid-based cohesion distance computation is $O(N)$ for a given set of N images.

Link-based cohesion. A common link-based cohesion measure is to derive a value from the pairwise distances among data points in a given dataset. However, the computational cost is $O(N^2)$ for a given collection of N data points. In our work, we adopt the cohesion measure proposed in [7] with certain modifications. The cohesion of a given set of data points is the proportion of "coherent" pairs among all pairs in the set. The "coherent" pairs are determined by a binary function $\delta(i_j, i_k)$, shown in Equation 2, where τ is a predetermined threshold.

$$\delta(i_j, i_k) = \begin{cases} 1 & \text{if } dist(i_j, i_k) \leq \tau \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

A tag may be used to annotate a large number of images, and pair-wise distance computation is computational costly. Therefore, a fixed number of pairs (say M) are randomly sampled from all pairs and the cohesion measure is the proportion of "coherent" pairs among the M sampled pairs. Let i_j^ℓ and i_k^ℓ ($k \neq j$) be the ℓ-th ($1 \leq \ell \leq M$) sampled pair from I_t. Then, the link-based cohesion distance, denoted by $\Phi_{link}(I_t, I_t)$, is given by the following equation.

$$\Phi_{link}(I_t, I_t) = \frac{1}{M}\sum_{i_j^\ell, i_k^\ell \in I_t} \delta(i_j^\ell, i_k^\ell) \quad (3)$$

In our experiments, we set $M = 10{,}000$. In order to determine the threshold τ, a larger number of pairs (20,000 in our experiments) are randomly sampled from the collection I, and τ is set to $\tau = \mu - 2\sigma$ where μ and σ are the mean and standard deviation of the sampled pairs, respectively. The computational cost is $O(1)$ for all tags for a given M.

3.2.2 Separation Distance Computation

We adopt the centroid-based and clarity-based measures to quantify the separation distance between I_t and I.

Centroid-based separation. The separation distance, denoted by $\Psi_{cent}(I_t, I)$, is the distance between the centroids of I_t and I respectively, given by the following equation.

$$\Psi_{cent}(I_t, I) = dist(Cent(I_t), Cent(I)) \quad (4)$$

Note that the computational cost of the above equation is $O(N)$ for a given set of N images.

Clarity-based separation. With bag of visual-words representation, an image can be treated as a document except that visual words in images are codewords rather than a unit of language. The clarity-based separation distance between I_t and I is the KL-divergence between *tag language model* $P(w|I_t)$, and *collection language model* $p(w|I)$ [4], as given by Equation 5. This distance is also known as the *clarity score* of tag t_v in this paper as it is defined similar to the query clarity score in [4] with certain modifications[3].

$$\Psi_{clar}(I_t, I) = \sum_w P(w|I_t)\log_2\frac{P(w|I_t)}{P(w|I)} \quad (5)$$

In the above equation, $P(w|I)$ is estimated by the relative visual-word frequency in the collection. Assume that every image in I_t has equal chance of being observed[4]. That is, $P(i|I_t) = 1/|I_t|$. Then, $P(w|I_t)$ is estimated using Equation 6 where $P_{ml}(w|i)$ is the maximum likelihood of observing a visual-word w in image i.

$$P_{ml}(w|I_t) = \frac{1}{|I_t|}\sum_{i \in I_t} P_{ml}(w|i) \quad (6)$$

The estimated tag language model is further smoothed using Jelinek-Mercer smoothing in Equation 7 with $\lambda = 0.99$ in our experiments[5]. Figure 3 plots the tag language models estimated for tags sunset and Asia using the bag of visual-word features provided in the NUS-WIDE dataset[6] (see Section 4.1 for more details of the dataset). Clearly, the language model of tag sunset is much more divergent from the collection language model than the language model of tag Asia.

$$P(w|I_t) = \lambda P_{ml}(w|I_t) + (1-\lambda)P(w|I) \quad (7)$$

The time complexity of clarity-based separation is $O(N)$ for a tag used to annotate N images.

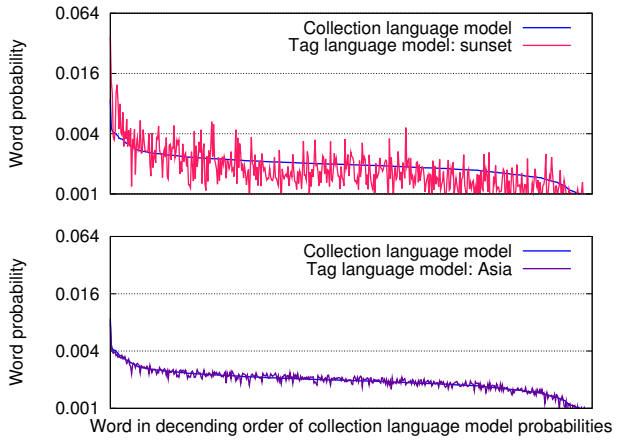


Figure 3: Tag language models for sunset and Asia derived from 500D bag of visual-words representation. The language model for the visual-representative tag sunset is much divergent from the collection language model. The language model for tag Asia, however, is similar to the collection language model.

[2] Cosine distance is derived by 1-cosine similarity.

[3] In [4], query clarity score is computed through top-500 ranked documents returned by a retrieval model. In our setting, the relevance between an image to a tag is unknown and a boolean retrieval model is adopted (see Equation 6).

[4] In [17], the distance between an image to the centroid of I_t was considered in estimating $P(w|I_t)$. However, in our experiments, we observed that very similar results were obtained using the simple estimation $P(i|I_t) = 1/|I_t|$. We hence adopt the latter for more efficient computation.

[5] Observe that a relatively large λ is set in our experiments as we are more interested in the difference (or separation) between the two probability distributions.

[6] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

3.3 Distance Normalization

In social tagging environment, it is well-known that some tags are more frequently used than others. Let *tag frequency* refers to the number of images annotated by a tag t, i.e., $|I_t|$. Figure 4 illustrates the tag frequency distribution in the NUS-WIDE dataset containing 269,648 images from *Flickr* [3]. Observe that tag frequency follows a power law distribution. With the above characteristics in mind, is it necessary to normalize the two distance measures in order to incorporate the effect of tag frequency distribution? In this section, we address this issue.

Cohesion distance is not affected by tag frequency. In general, more images tagged by t_v means a more accurate estimation of Φ_{cent}. For link-based cohesion, the same M number of pairs are sampled and the same threshold τ is applied for all tags. Hence, Φ_{link} is independent of tag frequencies with the condition that for every tag, M is smaller than the number of distinct pairs in I_t.

Now consider the separation distance. Both centroid- and clarity-based measures are affected by tag frequency. Recall that a tag t_v is considered visually representative if $SeD(I_t, I) \geq SeD(I_{rand}, I)$ where $|I_{rand}| = |I_t|$ and I_{rand} is randomly sampled from I. For random sampling, naturally, the larger the size of I_{rand}, the smaller is the distance $SeD(I_{rand}, I)$. Consider the extreme case, when the size of I_{rand} approaches the size of I, $SeD(I_{rand}, I)$ approaches to 0. Tags with different frequencies are therefore compared against different $SeD(I_{rand}, I)$'s. In the sequel, we use clarity-based separation to illustrate the impact of tag frequency.

Let t_d be a dummy tag assigned to every image in I_{rand}. We can then compute the clarity score of t_d using Equation 5. Figure 5 shows the average clarity scores and the standard deviations derived from 500 dummy tags with respect to each tag frequency on the x-axis. It demonstrates that the average clarity scores and standard deviations of dummy tags decrease as expected with the increase of tag frequency. Similar trend can also be observed for the centroid-based separation measure, which is however not shown for the interest of page limit.

Let I_d be the set of images tagged by a dummy tag t_d. Let $\mu(t_d)$ and $\sigma(t_d)$ be the expected tag clarity score and standard deviation derived from dummy tags having same tag frequency of a given tag t_v (i.e., $|I_t| = |I_d|$). We applied zero-mean normalization to derive $\Psi_{norm}(I_t, I)$ in Equation 8. The normalization is applied to distances obtained from Equations 4 and 5 (denoted by Ψ below) with $\mu(t_d)$ and $\sigma(t_d)$ are derived using the corresponding distance definitions.

$$\Psi_{norm}(I_t, I) = \frac{\Psi - \mu(t_d)}{\sigma(t_d)} \quad (8)$$

To minimize the computation cost, instead of computing $\mu(t_d)$ and $\sigma(t_d)$ for every tag frequency, we binned the frequencies with varying bin sizes. The first bin covers tag frequency from b_0 =100 to b_1 =110. Then $b_{n+1} = (1+10\%) \times b_n$ ($n \geq 0$) until the last bin covers the tag with highest frequency in our dataset. For each bin starting with b_n, 500 dummy tags with tag frequencies randomly sampled within $[b_n, b_{n+1})$ are used to derive the expected distance and standard deviation. Based on this framework, each separation distance $\Psi(I_t, I)$ of a tag t_v is normalized using dummy tags generated with frequencies within 10% of variation from its frequency $|I_t|$.

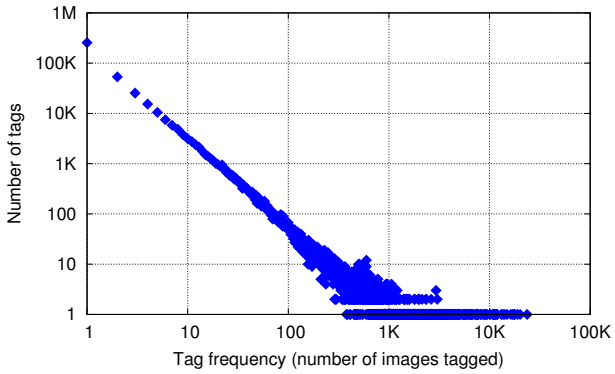


Figure 4: Tag frequency in the NUS-WIDE dataset follows a power-law distribution.

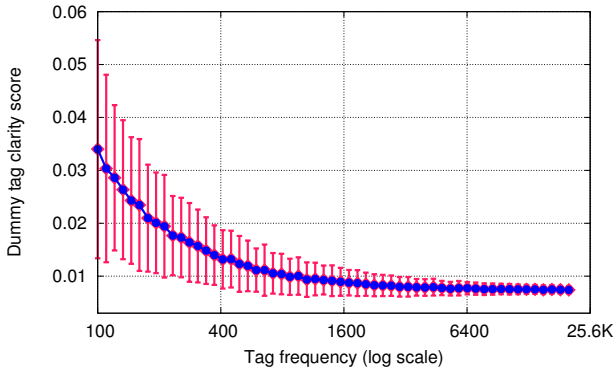


Figure 5: Average clarity scores and the standard deviations (stddev) derived from 500 dummy tags for each frequency on the x-axis. Both clarity scores and stddev values decrease with the increase of tag frequency.

4. EXPERIMENTS

4.1 Dataset

We used the NUS-WIDE dataset containing 269,648 images from *Flickr* [3]. In total six types of low-level features are provided for images in the dataset including global features such as color, edge, texture, and local feature of bag of visual-words. In particular, for our experiments we evaluated the following global and local features separately.

- **Global features.** We used three types of global features, including 64-D color histogram, 73-D edge direction histogram, and 128-D wavelet texture features. More details of these features are found in [3]. The three types of features for each image were aggregated into a 265-D vector after unit-length normalization on each type of features. We used Euclidian distance to compute distances between images or centroids.

- **Local features.** We used the 500-D bag of visual-words. Images are processed very much like documents in this setting and we adopted $tf \times idf$ word weighting scheme and cosine similarity for distance computation. For clarity-based measures, the language models were estimated on the 500 visual-words.

Observe that in our dataset more than $424K$ unique tags each appears at least once (see Figure 4). On average, each image is annotated with 18 tags. In our experiments, we mainly focus on the frequently-used tags such that each tag has been used to tag at least

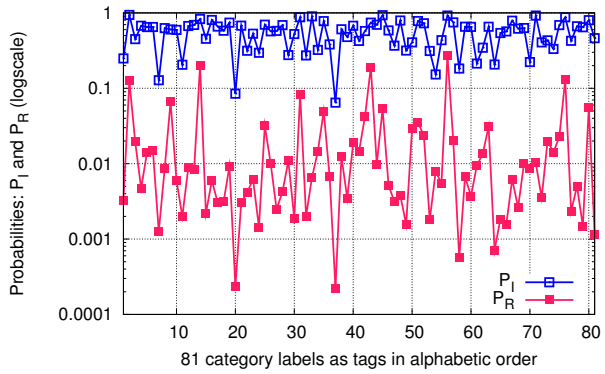


Figure 6: For all the 81 visual concepts, the probabilities of observing the visual concepts among its tagged images P_I are significantly larger than the probabilities of observing the visual concepts among all images in the dataset P_R.

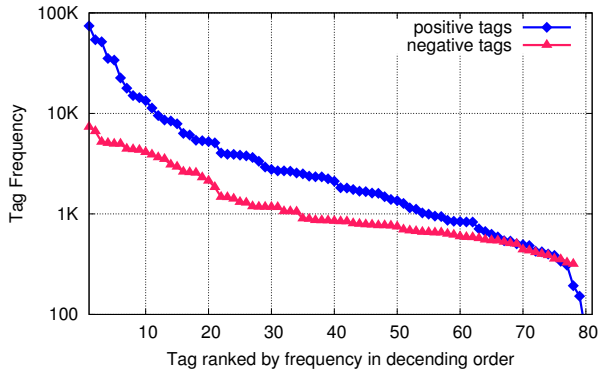


Figure 7: Frequencies of positive and negative tags. Note that the positive tags happen to be much more frequently-used than negative tags, leading to a skewed frequency distribution.

Table 2: The seven distance measures (methods).

Method	Distance	*Dist*	Feature	Equation
CCentL	Cohesion	Centroid	local	Eq. 1
CCentG	Cohesion	Centroid	global	Eq. 1
CLinkL	Cohesion	Link	local	Eq. 3
CLinkG	Cohesion	Link	global	Eq. 3
SCentL	Separation	Centroid	local	Eq. 4 and 8
SCentG	Separation	Centroid	global	Eq. 4 and 8
SClarL	Separation	Clarity	local	Eq. 5 and 8

0.1% (or 270) images in the dataset. There are 2568 such frequent tags, which are also known as *popular tags* in this paper.

Besides the low-level features and tags, images in NUS-WIDE dataset are manually assigned to a pre-defined list of 81 categories including 31 categories for object and 33 for scene. Interestingly, all the category labels also appear as tags. Note that the manual annotations provide us with the ground-truth of images relevant to each of the 81 visual concepts. That is, $\mathcal{R}_v$ is provided by the dataset for each of the 81 visual concepts.

4.2 Verification of Assumptions

Recall from Section 3.1, our proposed technique is based on two assumptions. We now empirically verify the first assumption here. Note that it is not necessary to verify the second assumption as it is based on a large body of literature in CBIR [5, 14].

The first assumption states that if a tag well describes a visual concept, then the probability of observing the visual concept among its tagged images is larger than the probability of observing it among all images. Let $P_I = \frac{|I_t \cap \mathcal{R}_v|}{|I_t|}$ and $P_R = \frac{|\mathcal{R}_v|}{|I|}$ denote the two probabilities, respectively. As the images in the dataset were manually assigned to 81 categories and the 81 category labels also appeared as tags, we have both I_t (tagged by users) and $\mathcal{R}_v$ (assigned by experts) for these 81 tags.

Figure 6 plots the curves for P_I and P_R (all 81 tags). The figure clearly states that $P_I \gg P_R$ for all tags. On average, $P_I = 0.55$ and $P_R = 0.02$. The three tags with the largest P_I's are all objects: 0.94 for animal, 0.93 for plants, and 0.92 for toy. The three tags with smallest P_I's are map, earthquake, and book. Nevertheless, all the latter three tags are relatively less popular in the dataset with tag frequencies of 372, 566, and 766 respectively among more than 269K images. Hence, our first assumption holds in this dataset.

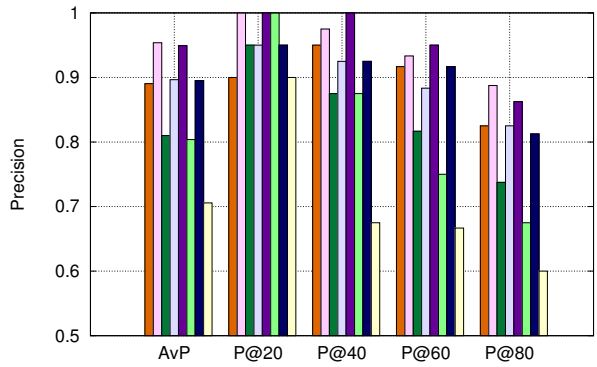
4.3 Methods and Evaluation Metric

In this section, we evaluate the effectiveness of cohesion and separation distances of identifying visually representative tags. Depending on the type of distances (see Sections 3.2 and 3.3) and the type of low-level features, we investigate seven distance measures (or methods) as listed in Table 2. For a given dataset, each of the seven methods will return a list of tags ranked according to the method-specific distance definition and feature space. For presentation clarity, we standardize all ranking lists such that the top-ranked tags in each list are meant to be more visually representative than bottom-ranked tags.

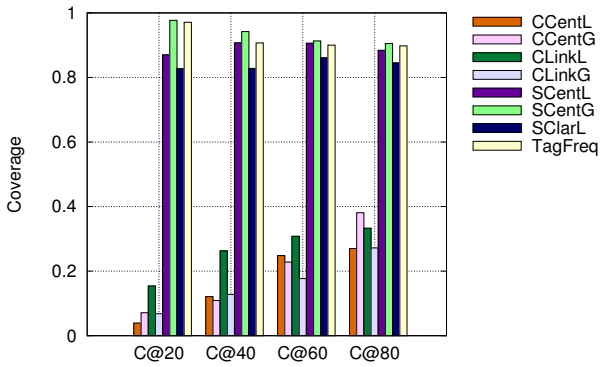
Average Precision and Precision@N. Given two ranking lists, the evaluation of the effectiveness of the two measures is non-trivial. First of all, it is hard to determine a ground truth ranking. For example, among tags that are visually representative, it is hard to determine their relative order purely based on their visual-representativeness, e.g., sunset, zebra, and architecture. Therefore, it is more reasonable to evaluate a partial order such that the visually representative tags (or *positive* tags) are ranked higher than tags that are not visually representative (or *negative* tags). For this purpose, we adopt two measures, namely, *Average Precision* (denoted by *AvP*) and *Precision@N* (denoted by *P@N*). P@N is the precision obtained among the top-N ranked tags. We report multiple P@N's for different N's depending on the number of tags in the ranking. Average precision is the average of precisions obtained at the point of each of the positive tags in the ranking.

Both *AvP* and *P@N* reflect how effectively a method rank visually representative tags higher than those that are not visually representative. The two measures, however, may not give a complete picture of each method.

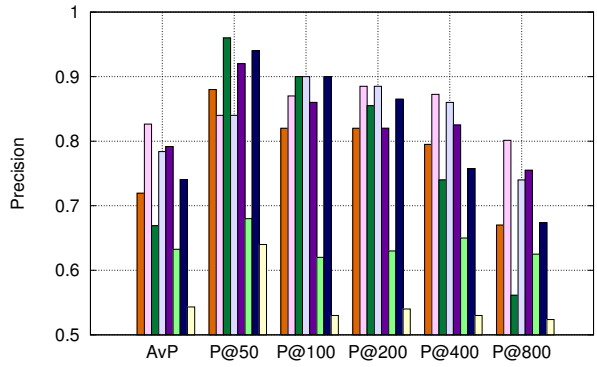
Coverage@N. Consider a tag recommendation scenario where only the *visual content* of images are available. Hence only visually representative tags may be recommended. The pool of tags to be recommended are taken from the ranking by each of the seven methods above. The method that ranks frequently-used visually representative tags higher is more favorable, as these tags can effectively annotate more images. At the same time, a method that mistakenly ranks a frequently-used but not visually representative tag higher would potentially affect the performance of tag recommendation adversely. For this reason, we propose a metric called **Coverage**@N (denoted by *C@N*). Informally, *C@N* is a variant of the widely adopted Normalized Discounted Cumulative Gain



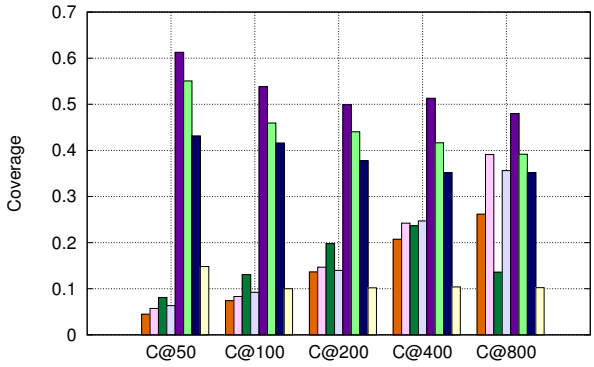
(a) Average Precision and Precision@N on 81 category labels



(b) Coverage@N on 81 category labels



(c) Average Precision and Precision@N on 1576 popular tags



(d) Coverage@N on 1576 popular tags

Figure 8: Performance on the 81 category labels as tags ((a) and (b)) and the popular tags ((c) and (d)). From the results, both *CCentG* and *SCentL* achieve very good average precision and precision@N. However, all four cohesion-based methods perform poorly on coverage measure as their top-ranked tags are relatively less popular. Overall *SCentL* is the most superior method w.r.t both precision and coverage measures.

(NDCG) measure in Information Retrieval [9]. Given a tag ranking $[t_1, t_2, \ldots, t_m]$, $C@N$ $(1 \leq N \leq m)$ is given by the following equation where r is the ranking position, $\rho(t_r)$ is the popularity of tag t_r, and $\upsilon(t_r)$ is a weighting function.

$$C@N = \frac{1}{Z} \sum_{r=1}^{N} \frac{\rho(t_r) \times \upsilon(t_r)}{log(r+1)} \quad (9)$$

We set $\rho(t_r)$ to be its tag frequency; $\upsilon(t_r) = 1$ for a visually representative tag and $\upsilon(t_r) = -1$ for a tag that is not visually representative. Note that similar to NDCG, Z is a normalization factor such that a perfect ranking[7] at N will give $C@N$ of 1 (see [9] for more details on Z).

4.4 Evaluation with Partial Ground-truth

Our first set of experiments is to evaluate the effectiveness of the seven methods of identifying the visually representative tags, with partial ground-truth. We first identify the positive and negative tags following our discussion in Section 4.3.

Positive tags. Recall that images in the dataset were manually assigned to 81 categories. Hence, for these 81 tags (also as category labels), $\mathcal{R}_v$ is available for computing cohesion and separation distances. These 81 tags are positive tags as images of the same category share the same visual concept described by the corresponding category label.

Table 3: Tag category labels and example tags in each category.

Label		#tags	Examples
o	object	442	zebra, architecture, girl
s	scene	281	sky, nature, sunset
a	activity	45	travel, dancing, wedding
c	color	28	red, blue, blackandwhite
u	picture type	18	wideangle, macro, hdr
l	location	273	asia, china, europe
r	self-reference	217	deleteme, 10faves, selfportrait
n	opinion	157	beautiful, colorful, supershot
m	camera	89	canon, nikon, 400d
t	time	26	2008, october, spring

Negative tags. Tags related to time and location like county names often have little relevance to visual concepts. Among the 2568 popular tags, we selected 17 time tags (e.g., year 2004 - 2008, month January - December) and 61 location tags for continent and country names (e.g., Europe, Japan). Note that, we do not use tags for very specific locations (e.g., a small town or a park) as negative tags. The reason is that many images tagged with a very specific location may be about the landmarks of that location and hence the tags become visually representative of those landmarks.

[7] More frequently used positive tags shall be ranked higher than less frequently used positive tags.

Table 4: The most (t1-15) and least (b1-15) visually representative tags identified by the seven methods. Tags that had been used as visual concepts in eariler work are highlighted in *bold* and underlined.

	Cohesion distance				Separation distance		
Rank	*CCentL*	*CCentG*	*CLinkL*	*CLinkG*	*SCentL*	*SCentG*	*SClarL*
t1	o.bigcats	o.motorsport	s.thunderstorm	o.whitetail	s.***sunset***	s.***sky***	s.***sunset***
t2	o.wolves	o.pandas	s.lightning	o.pandas	s.***sky***	c.blue	s.fog
t3	o.bigcat	s.seabaths	o.leopard	s.seabaths	s.***clouds***	s.***water***	s.***sky***
t4	s.nationalzoo	s.oceanbaths	o.whitetail	s.oceanbaths	s.***landscape***	u.hdr	s.silhouette
t5	o.whitetail	o.whitetail	s.thunder	o.motorsport	s.***night***	c.green	s.sunrise
t6	o.pandas	o.bigcat	o.buck	o.wolves	s.***sea***	o.***architecture***	u.charts
t7	o.***tiger***	c.bwdreams	o.bigcats	o.en.species	s.sunrise	s.***sea***	o.***sun***
t8	s.thunderstorm	o.cruiseship	o.***tiger***	n.bwdreams	s.fog	s.***nature***	s.mist
t9	o.lions	o.wolves	o.***zebra***	o.cruiseship	u.hdr	c.red	s.***sea***
t10	o.buck	s.lichen	s.storms	o.bigcat	s.silhouette	s.***night***	s.***clouds***
t11	o.panda	o.buck	s.foggy	s.lichen	s.***beach***	s.***clouds***	s.lightning
t12	o.en.species	r.theface	s.fog	o.buck	o.***sun***	s.***landscape***	s.***beach***
t13	s.lightning	a.carracing	o.pinhole	c.blackwhite	c.blue	n.anawesomeshot	s.***landscape***
t14	o.cub	u.sketches	o.bigcat	o.panda	s.***lake***	n.aplusphoto	s.dunes
t15	o.***lion***	n.masterphotos	o.wolves	c.blackandwhite	m.longexposure	s.***sunset***	c.blue
b1	s.ceiling	r.blog	o.clothes	l.seattle	r.100views	n.flickrexplore	o.***people***
b2	o.curves	l.mexico	s.conf.room	t.2007	t.sunday	t.sunday	c.brown
b3	u.charts	n.cool	o.bag	o.toys	t.february	n.large	l.asia
b4	o.circle	m.photoshop	o.***computer***	o.***sign***	r.pics	r.pics	l.japan
b5	n.symmetry	c.colour	s.gym	o.baseball	n.huge	t.december	l.france
b6	o.officebuilding	r.photos	a.work	o.cellphones	n.flickrexplore	r.saveme3	l.washington
b7	n.creative	l.wisconsin	r.auto	s.museum	n.passion	r.saveme6	t.2008
b8	o.curve	l.texas	s.***kitchen***	l.taiwan	n.pictureperfect	n.romance	l.china
b9	o.card	l.maryland	o.***tattoo***	o.cables	r.is	r.saveme2	r.photograph
b10	o.architektur	l.seattle	r.individual	n.cool	r.picnik	r.blogged	t.july
b11	o.skyscraper	l.florida	s.diningroom	o.pavilions	r.photos	r.childhood	r.picture
b12	l.munich	n.fabulous	o.boots	r.thecontinuum	r.ruby.p	s.down	l.virginia
b13	o.glass	r.showpixels	o.cutout	t.2006	n.ilovemypics	r.deleteme9	l.india
b14	n.artlegacy	r.the	r.2	a.jump	t.august	r.saveme4	l.ohio
b15	o.cables	r.geotagged	o.computers	r.3	r.3	r.save5	t.august

abbreviations: showpixels (showmeyourqualitypixels), en.species (endangeredspecies), conf.(conference), ruby.p (rubyphotographer)

***TagFreq* as baseline**. Recall from Section 2, the most relevant work to our proposed approach that we can compare empirically is [15]. Unfortunately, despite our best efforts (including contacting the authors), due to legal restrictions we could not get the source code of [15]. Hence, we compare our proposed approach with *TagFreq* which has been adopted as the baseline in [10] for image annotation. *TagFreq* ranks tags simply by their frequencies in descending order. In this experiments, since positive tags happen to be much more frequently-used than negative tags (see Figure 7), a simple rank by frequency has the potential to achieve good precision and coverage as the top-ranked tags by frequency are more likely to be positive tags.

Results. Figures 8(a) and 8(b) depict the results of our study. Both *CCentG* and *SCentL* achieve surprisingly good *AvP* (higher than 0.95) and *TagFreq* shows the worst results with *AvP* of 0.7. For *P@20–P@80*, all seven methods follow similar trend, in consistent with AvP. Figure 8(b) shows the coverage measure of all seven methods. As illustrated, the three separation-based methods favor frequently-used tags. Although the four cohesion-based methods achieve fairly good precision, they perform poorly on coverage because their top-ranked tags are relatively less popular. Lastly, *TagFreq* shows good coverage due to skewed frequency distribution among positive and negative tags (see Figure 7).

4.5 Evaluation on Popular Tags

Popular tags. Among the 2568 popular tags identified in Section 4.1, we manually labeled 1576 tags into 10 categories listed in Table 3 where each labeled tag has a relatively clear semantic meaning. Among the 10 categories, object, scene, location, self-reference, and emotion have more number of tags which is consistent with that reported in [1]. In general, tags fall into the top 5 categories in Table 3 are believed to be more descriptive of the image contents than the tags in the bottom 5 categories. In this set of experiments, we consider the 814 tags in the top 5 categories as positive tags and the remaining 762 tags in the bottom 5 categories as negative tags. To partially verify our labeling, we collected the category labels/tags used as visual concepts in earlier work [3, 8, 12, 14] for image classification, annotation, and other tasks. We found that among the 814 positive tags, 118 distinct tags had been used in earlier work as visual concepts and none of the visual concepts in earlier work match the 762 negative tags.

Precision and coverage. Figures 8(c) and 8(d) plot the precision and coverage of all the methods. As illustrated, *CCentG* and *SCentL* demonstrate the best and second best precisions, respectively. Such results are consistent with the results of the first set of experiments. Also, as illustrated in Figure 8(d), the three separation-based methods perform significantly better than the four

cohesion-based methods on coverage. Note that the *TagFreq* method lose out significantly on both precision and coverage due to more balanced frequency distribution between positive and negative tags in this experiment. Overall, *SCentL* is the most superior method *w.r.t* both precision and coverage.

Top and bottom ranked tags. Table 4 lists the top ($t1$-$t15$) and bottom ($b1$-$b15$) ranked tags based on the seven methods. All tags are prefixed with their label categories referencing Table 3. Among the top-15 tags, it is interesting to observe that separation-based methods favor tags of type scene, and cohesion-based methods favor object tags. More importantly, for cohesion-based methods, many top ranked object tags refer to very specific visual concepts, such as wolves, bigcats, and panda. These tags, however, are not very frequently used compared to others, which explains the lower coverage by the four methods. In contrasts, separation-based methods rank more scene tags at the top, and many of these tags are extremely popular (e.g., sky, sunset, and sea). These popular tags contribute to the high coverage of the three methods. Comparing global and local features, *SCentG* ranks more color tags (e.g., blue, green, and red) on the top than *SCentL* and *SClarL*. Among the bottom-ranked 15 tags, more tags under location, emotion and self-reference categories are identified by separation-based methods whereas more object tags are identified by cohesion-based methods.

Recall that among the 814 positive tags, 118 tags had been used as visual concepts in earlier work. Tags matching these 118 tags are highlighted in ***bold*** in Table 4. Both *SCentL* and *SCentG* match 9 among the top 15 tags, followed by *SClarL* with 7 hits. As the 118 tags were all manually and carefully selected in earlier work, our experimental results support our claim that the proposed technique could serve as guideline on the selection of visual concepts.

In summary, separation-based methods are more effective in identifying frequently-used visually representative tags. In particular, both sets of experiments demonstrate that *SCentL* is the most effective method.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel technique to quantify tag visual-representativeness for social images. Tag visual-representativeness reflects the consensus implicitly developed among users on image annotation which may lead to better understanding of tag usage. Specifically, we bridge the two orthogonal areas of social image tagging and query performance prediction in Web search, to propose two simple and yet effective metrics, namely cohesion and separation, to quantify tag visual-representativeness. Plugged in with different distance definitions for the chosen image low-level feature representations, our empirical study demonstrated that separation-based metrics were the most effective in identifying frequently-used tags that are representative of visual concepts. More importantly, all proposed measures are efficient to compute.

There are at least two interesting directions worth further exploration. First, it is shown in our experiments, the cohesion and separation measures favor different types of tags (e.g., scene and object) with different popularity levels. This calls for an aggregated measure to take advantage of various cohesion and separation measures for more accurate quantification of the visual-representativeness of tags. Combining some of the proposed measures to achieve a more accurate visual-representative measure naturally becomes part of future work. Second, the proposed measures compute a visual-representativeness score for a tag using all its annotated images. Hence these measures fail to distinguish tags with more than one visual concepts. For example, jaguar could be representative for two distinct visual concepts, namely car and animal. As part of future work, we wish to investigate techniques to identify tags representing multiple visual concepts. Other than these two directions, the evaluation of visual-representativeness is also an interesting research topic.

6. REFERENCES

[1] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *Proc. CIKM*, 2008.

[2] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. What makes a query difficult? In *Proc. SIGIR*, 2006.

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

[4] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proc. SIGIR*, 2002.

[5] 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.

[6] C. Hauff, V. Murdock, and R. Baeza-Yates. Improved query difficulty prediction for the web. In *Proc. of CIKM*, 2008.

[7] J. He, W. Weerkamp, M. Larson, and M. de Rijke. An effective coherence measure to determine topical consistency in user-generated content. *IJDAR*, 12(3):185–203, 2009.

[8] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *Proc. ACM MIR*, 2008.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[10] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):985–1002, 2008.

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

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

[13] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proc. WWW*, 2009.

[14] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40:262 – 282, 2007.

[15] Y. Lu, L. Zhang, Q. Tian, and W.-Y. Ma. What are the high-level concepts with small semantic gaps? In *Proc. CVPR*, 2008.

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

[17] A. Sun and S. S. Bhowmick. Image tag clarity: In search of visual-representative tags for social images. In *Proc. of ACM SIGMM Workshop on Social Media (WSM)*, 2009.

[18] A. Sun and A. Datta. On stability, clarity, and co-occurrence of self-tagging. In *Proc. of WSDM (Late Breaking-Results)*, 2009.

[19] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2005.

[20] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *Proc. ACM MM*, 2009.

[21] J. Teevan, S. T. Dumais, and D. J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proc. SIGIR*, 2008.

[22] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *Proc. WWW*, 2009.

[23] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma. Annotating images by mining image search results. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1919 – 1932, 2008.

[24] K. Weinberger, M. Slaney, and R. van Zwol. Resolving tag ambiguity. In *Proc. ACM MM*, 2008.

[25] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *Proc. of MM'08*, pages 31–40, Vancouver, Canada, 2008.

[26] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *Proc. SIGIR*, 2007.