Image Tag Clarity: In Search of Visual-Representative Tags for 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

Tags associated with images in various social media sharing web sites are valuable information source for superior image retrieval experiences. Due to the nature of tagging, many tags associated with images are not visually descriptive. In this paper, we propose *Normalized Image Tag Clarity* (NITC) to evaluate the effectiveness of a tag in describing the visual content of its annotated images. It is measured by computing the zero-mean normalized distance between the *tag language model* estimated from the images annotated by the tag and the *collection language model*. The visual-representative tags that are commonly used to annotate visually similar images are given high tag clarity scores. Evaluated on a large real-world dataset containing more than 269K images and their associated tags, we show that NITC score can effectively identify the visual-representative tags from all tags contributed by users. We also demonstrate through experiments that most *popular* tags are indeed visually representative.

Categories and Subject Descriptors

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

Keywords

Image tag clarity, Visual-representative tag, Flickr, Language model.

1. INTRODUCTION

With the advances in digital photography (e.g., digital cameras and mobile phones) and social media sharing web sites, a huge number of multimedia content is now available online. Most of these sites enable users to annotate web objects including images with free tags (e.g., aircraft, lake, sky). For instance, most images accessible through Flickr[1] are annotated with tags from their uploaders as well as other users. A key consequence of the availability of such tags as meta-data is that it has significantly facilitated web image search and organization as this rich collection of tags provides more information than we can possibly extract from content-based algorithms.

Due to the popularity of tags, there have been increasing research efforts to better understand and exploit tag usage patterns for information retrieval and other related tasks. One such effort is to make better use of the tags associated with images for superior image retrieval experiences. However, this is still a challenging research problem as it is well known that tags are noisy and imprecise [1]. As discussed in [4], tags are more for personal use than others' benefit. Consequently, two similar images may be associated with significantly different sets of tags from different users, especially when images can only be annotated by users with *tagging permissions* (e.g., in Flickr, only the uploader and his/her contacts can tag an image). Further, tags associated with an image may describe the image from significantly different perspectives. For example, consider a photo uploaded by Sally which she took using her Canon 40D camera at Sentosa when she traveled to Singapore in 2008. This image may be annotated by different tags such as Canon, 40D, 2008, Singapore, travel, beach, Sentosa, and many others. Notice that tags like 2008 and Canon do not effectively describe the visual content of the image. Consequently, these tags maybe considered as noise in many applications. As the presence of such noise may reduce the usefulness of tags in image retrieval, "de-noising" tags has been recently identified as one of the key research challenges in [1]. Such de-noising of tags also enables us to build more effective tag ranking and recommendation services [7].

In this paper, we take a step towards addressing the above challenge. We focus on identifying *visual-representative* tags from all tags assigned to images so that less representative tags can be eliminated. Intuitively, a tag is *visual-representative* if it effectively describes the *visual content* of the images. A visual-representative tag (such as sky, sunset, and beach) easily suggests the scene an image may describe even before the image is presented to a user. On the other hand, tags like 2008 and Asia often fail to suggest anything

WSM'09, October 23, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-759-2/09/10 ...$10.00.

[1] http://www.flickr.com

meaningful related to the visual content of the annotated image.

We propose the notion of *Normalized Image Tag Clarity* (NITC) to identify visual-representative tags. It is inspired by the *clarity score* proposed for query performance prediction in ad-hoc information retrieval for textual documents [2]. Note that clarity score cannot be directly applied to annotated images as keywords of a query literally appears in the retrieved text documents whereas the tags associated with an image do not explicitly appear in it.

Our experimental study with the NUS-WIDE dataset [1], containing 269,648 images from Flickr, demonstrates that the proposed NITC measure can effectively identify and rank the visually representative tags. We also analyze the relationship between *tag popularity* and clarity and show that most popular tags are indeed visually representative. This suggests that users share common knowledge on the semantics of these tags and the visual content of the annotated images.

The rest of the paper is organized as follows. In Section 2, we review the related work with emphasis on clarity score for query performance prediction as well as image tagging. Section 3 discusses the notion of image tag clarity. The details of the dataset and experimental results are reported in Section 4. We conclude this paper in Section 5.

2. RELATED WORK

Recall that our proposed image tag clarity measure is inspired by the notion of clarity score proposed for query performance prediction in ad-hoc retrieval. Hence, we begin by reviewing the clarity score measure. Next, we discuss relevant research efforts in annotating web objects with tags.

2.1 Clarity Score

Query performance prediction is to predict the effectiveness of a keyword query in retrieving relevance documents from a document collection [2]. The prediction enables a search engine to answer poorly performing queries more effectively through alternative retrieval strategies (e.g., query expansion) [5, 11, 15, 16]. Depending on whether documents need to be retrieved for the query, the query performance prediction algorithms can be classified into two types: *pre-retrieval* and *post-retrieval* algorithms. Pre-retrieval algorithms rely on the statistics of the words in both the query and the collection. For instance, queries consisting of words with low document frequencies in the collection tend to perform better than queries with high document frequency words. Post-retrieval algorithms predict query performance based on the properties of the retrieved documents from the collection using the query. Among post-retrieval algorithms, one significant contribution is *clarity score* [2].

The *clarity score* of a query is computed as the *distance* between the *query language model* and the *collection language model.* If a query is effective in retrieving topically cohesive documents, then the query language model contains unusually large probabilities of words specific to the topic covered by the retrieved documents. Consequently, the distance between the query and the collection language models is large. If a query is ambiguous, then the documents covering various topics are likely to be retrieved. That is, the retrieved set of documents is similar to a set of documents through random sampling. As the word distribution in the retrieved documents is similar to that in the collection, the distance between them is small.

Formally, let Q be a query consisting of one or more query words $\{q|q \in Q\}$ and R be the set of top-K documents retrieved by Q from the collection $\mathcal{D}$. The value of K is predefined and set to 500 in [2]. Let w be an arbitrary word in the vocabulary. Then, the query language model $P(w|Q)$ is estimated by Equation 1, where $P(d|Q)$ is estimated using the Bayes' theorem as shown in Equation 2.

$$P(w|Q) = \sum_{d \in R} P(w|d)P(d|Q) \quad (1)$$

$$P(Q|d) = \prod_{q \in Q} P(q|d) \quad (2)$$

Observe that in both Equations 1 and 2, $P(w|d)$ and $P(q|d)$ is the relative frequency of word w (or q) in the document d linearly smoothed by w's relative frequency in the collection. The collection language model, $P(w|\mathcal{D})$, is estimated by the relative frequency of w in $\mathcal{D}$. Then, the clarity score of Q is the Kullback-Leibler (KL)-divergence between $P(w|Q)$ and $P(w|\mathcal{D})$, and is given by the following equation.

$$KL(Q\|\mathcal{D}) = \sum_{w} P(w|Q) \log_2 \frac{P(w|Q)}{P(w|\mathcal{D})} \quad (3)$$

2.2 Blog Tag Clarity

Tagging is a popular technique for annotating objects on the web. In our previous work [10], we introduced the notion of *tag clarity* in the context of users behavior study in self-tagging systems, i.e., blogs. The clarity score of a tag is defined by the KL-divergence between the tag language model (estimated from the blog posts associated with the tag) and the collection language model from all blog posts. As blogs are self-tagging, i.e., only the blogger could annotate his/her blog posts, the tag clarity was proposed to study whether users implicitly develop consensus on the semantic of the tags. We observed that frequently used tags are topic discriminative. This finding is partially consistent with the findings in this proposed work although the object (text vs image) of annotation and tagging rights (self-tagging vs permission-based tagging) are different.

Our proposed image tag clarity differs from the tag clarity in [10] in the following ways. We use the centrality document model for estimating the tag language model and the zero-mean normalization for deriving *normalized* image tag clarity. Both techniques were not used in [10]. As we shall see in Section 3.2, the normalization is a critical step in deriving more meaningful tag clarity scores and also give the scores a statistical meaning.

2.3 Tagging Images

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 [1]. Only few work exploit the relationship between a tag to the content of its annotated images. For a given image and its annotated tags, the *relevance* between the image and each tag is estimated through kernel density estimation in [7] and through k-nearest neighbor voting in [6]. In simple words, a tag is relevant to an image I if the tag has been used to annotate many images similar to I. The relevance score for a tag is therefore image-specific whereas in our case, the tag clarity score is *global.* For a given tag, the score reflects its effectiveness in visually describing all its annotated images. In this context, our work is also related to [9] 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 [9], for a given image I, 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. In contrast, our work differs in the following ways: (i) the computation of clarity score of a tag is purely based on its annotated images represented in visual space only; (ii) our task is to measure the visual-representativeness of a tag (i.e., a given concept) and not to mine concepts from textual descriptions; and (iii) our work does not rely on neighborhood relationships between images.

Very recently, *Flickr distance* was proposed to model two tags' similarity based on their annotated images [13]. 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 [13] in three aspects. First, our research objective is to measure the visual-representativeness of a single tag, not the relationship between tag pairs. Second, the language models are estimated from different image representations. Our language models are estimated on top of the widely adopted bag of visual words representation [8] while visual language model has it own definition in [13]. Third, we analyze the impact of tag frequency in its language modeling. In their work, a fixed number (i.e., 1000) of images for each tag were sampled for estimating its language model.

In [12], 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 tag clarity, the approach in [12] was purely based on tag co-occurrence without considering the content of annotated images.

3. IMAGE TAG CLARITY

Intuitively, a tag is visually representative if all the images annotated with the tag are visually similar to each other.

Our image tag clarity measure is based on the following framework. We consider a tag to be a keyword query and the set of images annotated with the tag are the retrieved documents based on a boolean retrieval model (which returns an image as long as the image is annotated with the tag with equal relevance score). Then the clarity score proposed for query performance prediction can be adopted to measure tag clarity if the visual content of the images can be represented by "word" vectors similar to that for representing textual documents. That is, if all images associated with the tag are visually similar, then the language model estimated from the set of retrieved images (or the tag language model) shall contain some "words" with unusually high probabilities specific to the tag making the distance between the tag and the collection language models large.

Among the various low-level features that are commonly used to represent images, *bag of visual words* feature represents images very much like textural documents [8]. In the sequel, we assume a bag of visual words is extracted to represent each image[2]. We also use "image" and "document" interchangeably because of this representation. We use d to denote an image.

3.1 Image Tag Clarity Score (ITC)

Let T be the set of images annotated by a tag t. Based on the clarity score definition in Equation 3, the *image tag clarity* score of t, denoted by $\textsc{itc}(t)$, is defined as the KL-divergence between the *tag language model* ($P(w|T)$) and the *collection language model* ($p(w|\mathcal{D})$). It is expressed by the following equation.

$$\textsc{itc}(t) = KL(T||\mathcal{D}) = \sum_w P(w|T)\log_2 \frac{P(w|T)}{P(w|\mathcal{D})} \qquad (4)$$

As a collection language model is often estimated by the relative word frequency in the collection, our main focus in this section is to estimate the tag language model $P(w|T)$. This is a challenging issue for the following reason. In textual documents, keywords in a query Q literally appears in the retrieved documents. Hence, the degree of relevance between a document and Q ($P(d|Q)$) can be estimated using Equation 2. However, in a bag of visual words representation, the tag and the words are from two different feature spaces. As a tag does not literally appear in images, the degree of relevance of an image to a tag is unknown. That is, $P(d|Q)$ in Equation 1 (or $P(d|T)$ in our setting) has to be estimated differently as Equation 2 cannot be directly applied.

Intuitively, there are at least two approaches to estimate the tag language model. First, we can simply treat all images equally representative of a tag t. Second, we can estimate the representativeness of images based on their distances to T's centroid. Images that are more close to the centroid of T are considered more representative and shall contribute more to the estimation of the tag language model.

The first approach estimates the tag language model as the average relative word frequency in the images with equal importance $\frac{1}{|T|}$. Hence, the tag language model, denoted by $P_s(w|T)$, is given by the following equation.

$$P_s(w|T) = \sum_{d\in T} \frac{1}{|T|} P_{ml}(w|d) \qquad (5)$$

Observe that it is consistent with the *small document model* used in [3] for blog feed search. Similar approach has also been used in modeling blog tag clarity in our earlier work [10].

In the second approach, also known as the *centrality document model*, the tag language model $P_c(w|T)$ is estimated using Equation 6, where $P(d|T)$ reflects the relative closeness of the image to T's centroid defined in Equation 7.

$$\begin{aligned} P_c(w|T) &= \sum_{d\in T} P_{ml}(w|d)P(d|T) && (6)\\ P(d|T) &= \frac{\varphi(d,T)}{\sum_{d\in T}\varphi(d,T)} && (7)\\ \varphi(d,T) &= \prod_{w\in d} P_s(w|T)^{P_{ml}(w|d)} && (8) \end{aligned}$$

In Equation 7, $\varphi(d,T)$ is a *centrality function* which defines the similarity between an image d to T. Let $P_s(w|T)$ be the tag language model estimated with small document model in

[2] Nevertheless, we believe that the image tag clarity score is generic and can be computed through other feature representations.

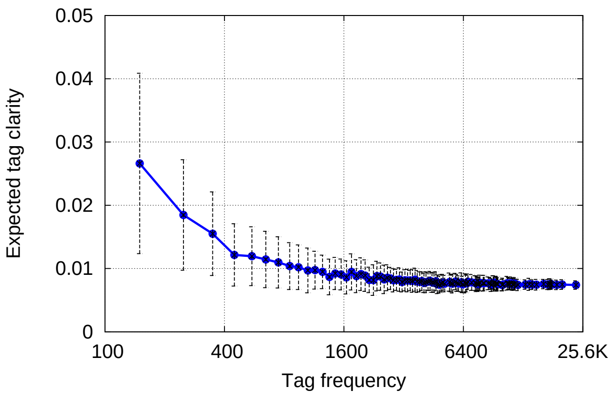


Figure 1: Expected tag clarity vs tag frequency.

Equation 5 and $P_{ml}(w|d)$ be the relative word frequency of w in image d. Then following [3], $\varphi(d,T)$ is defined to be the weighted geometric mean of word generation probabilities in T shown in Equation 8. The weight of each word is its likelihood in document d.

Intuitively the centrality document model better simulates the clarity score compared to the first approach. It also attempts to minimize the possible impact of outlier images in T as the images that are far from the centroid of T are considered less representative and contribute less in estimating the tag language model. Hence in this work, we adopt the centrality document model for estimating the tag language model.

The estimated tag language model is further smoothed using the Jelinek-Mercer smoothing with $\lambda = 0.99$.

$$P_{smoothed}(w|T) = \lambda P_c(w|T) + (1-\lambda)P(w|\mathcal{D}) \qquad (9)$$

3.2 Normalized Image Tag Clarity Score (NITC)

Recall from Section 2.1, the query language model is estimated from a fix number of top-K documents (e.g., K =500 in [2]). The clarity scores for all queries are therefore computed with the same number of documents. However in tagging, the tag distribution follows power-law distribution with a small set of tags much more frequently used than other tags (see Section 4.1). The sizes of T's for different tags can therefore be significantly different. We address this issue by *normalizing* the ITC score. We elaborate on this further.

Let us consider the task of assigning a tag to an image as a sampling process of picking up images from a large pool (i.e., collection $\mathcal{D}$). If the sampling is unbiased (i.e., uniform sampling), then the language model of the sampled images $P(w|T)$ naturally gets closer to $P(w|\mathcal{D})$ as T gets larger. Hence, the distance $KL(T\|\mathcal{D})$ becomes smaller. Assume that a tag t is assigned to an image d based on the visual content of the image. Then the sampling becomes biased as only images demonstrating some visual relevance to t are sampled. An image however may contain many objects or scenes. The objects contained in d but are not related to t become noise with respect to t. Hence, the chance of observing irrelevant words increases when T gets larger, again leading to a smaller $KL(T\|\mathcal{D})$. In summary, $KL(T\|\mathcal{D})$ may not accurately reflect the clarity of a tag as it is expected that $KL(T_1\|\mathcal{D}) < KL(T_2\|\mathcal{D})$ if $|T_1| > |T_2|$.

To overcome the impact of *tag frequency*, we applied zero-mean normalization to the image tag clarity scores. Tag frequency, denoted by $Freq(t)$, is the number of images a tag t is associated with in the given dataset, i.e., $|T|$. Let $\textsc{itc}(t)$ be the image tag clarity score of tag t as computed in Section 3.1. The *expected* image tag clarity score with respect to t is computed by randomly assigned dummy tags with the same frequency to images in the dataset. Let $\mu(Freq(t))$ and $\sigma(Freq(t))$ be the *expected tag clarity* and *standard deviation* obtained by assigning multiple dummy tags having the same frequency $Freq(t)$. Then, the *normalized image tag clarity* score, denoted by $\textsc{nitc}(t)$, is given by the following equation.

$$\textsc{nitc}(t) = \frac{\textsc{itc}(t) - \mu(Freq(t))}{\sigma(Freq(t))} \qquad (10)$$

Notice that the NITC score is in fact the number of standard deviations of a tag that is observed with respect to a randomly assigned dummy tag with the same tag frequency. For instance, if $\textsc{nitc}(t) \geq 2$, then the chance of a tag t being randomly assigned to images is smaller than 2.3%, assuming a normal distribution of the image tag clarity scores for tags with the same frequency. In this work, we assume that t is *visual-representative* (or representative in short) if $\textsc{nitc}(t) \geq 2$ and *highly representative* if $\textsc{nitc}(t) \geq 10$. Note that these threshold values are not fixed across all applications and can be adjusted according to a specific application.

To minimize the computation cost, instead of computing $\mu(Freq(t))$ and $\sigma(Freq(t))$ for every $Freq(t)$, we binned the tag frequencies with a bin size of 100. The expected tag clarity and standard deviation are computed for each bin with 100 dummy tags. A given tag clarity score is then normalized by $\mu(b)$ and $\sigma(b)$ where b is the bin $Freq(t)$ falls into. In our following discussions, all tag clarity scores refer to the normalized tag clarity scores.

Figure 1 plots the expected tag clarity scores and standard deviations against the binned tag frequencies computed from the NUS-WIDE dataset [1] (see Section 4.1 for more details of the dataset). Observe that the first point from the left plots the expected tag clarity score and the standard deviation for tags with frequencies within 100 and 200. Consistent with our earlier discussion, the expected clarity scores become smaller with the increase of tag frequency for all randomly assigned dummy tags. When the tag frequency is above 1000, the clarity scores become stable but the standard deviations continue to decrease as the tag frequency increases.

3.3 Time Complexity

The proposed tag language model can be estimated in $O(N)$ time for a tag associated with N images and requires at most three scans of the images (for computing Equations 6, 7, and 8). Note that the expected tag clarity scores need to be computed only once for a given dataset.

4. PERFORMANCE EVALUATION

In this section, we discuss the effectiveness of our proposed image tag clarity empirically. We begin by describing the dataset used for our experimental study.

4.1 Dataset

We used the NUS-WIDE dataset[3] containing 269,648 images

[3] `http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm` Accessed June 2009.

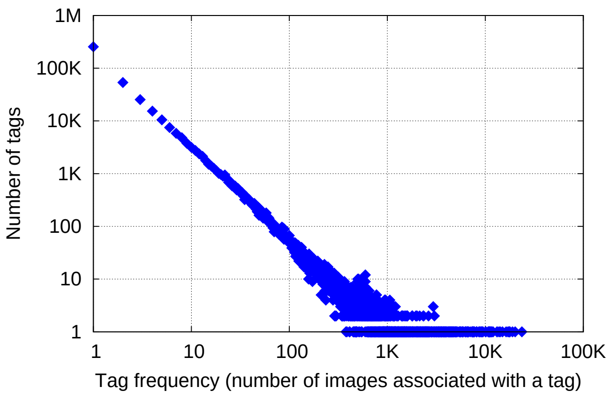


Figure 2: Tag frequency distribution.

from Flickr [1]. The images are assigned with zero, one or more categories (or concepts) from a pre-defined list of 81 categories. Among the 6 types of low-level features provided, we used the 500-D bag of visual words feature, so that each image is represented by a bag of words similar to a textual document. The original tag data without cleaning is used in this work. In the following, we report the tag distribution in the dataset.

The tag frequency distribution is reported in Figure 2. Similar to statistics related to many user-generated content, the tag frequency distribution follows a power-law distribution. In particular, there are more than 420K distinct tags appeared in the dataset and 5981 tags have been used to annotate at least 100 images each[4]. We consider these tags as *popular* tags and evaluate their clarity scores in the sequel.

4.2 Image Tag Clarity Evaluation

It is a challenging task to evaluate image tag clarity score even with users study. Given two tags, say sky and plane, it is difficult for a human annotator to conclude whether sky is more representative than plane. We therefore choose to evaluate the image tag clarity measure by the category labels provided by the NUS-WIDE dataset. We first report the NICT score distribution as well as the top-50 most and least representative tags identified through our experiments and give a detailed discussion on them.

The relationships between number of tags (tags are binned by $floor(\text{NITC}(t))$) with their NITC scores and the corresponding percentile are depicted in Figures 3(a) and 3(b), respectively. Observe that among the 5981 popular tags, 2372 tags (or about 40%) have NITC scores below 2 making them less representative. On the other hand, around 60% tags are identified as representative. Specifically, around 18% tags have NITC scores above 10. Consequently, these tags are considered highly representative. We believe that these percentages are fairly reasonable as significant number of tags are indeed representative for effective information organization and searching.

The top-50 most and least visual-representative tags are listed in Table 1 together with their NITC scores, frequencies, and frequency percentiles (denoted by $PerF(t)$). Observe that among the top-50 most representative tags, many of them describe common scenes (e.g., sunset, lightning, sea, and sky) or objects (e.g., zebra, jet, airplane and aircraft). As these are commonly used words, most users could easily pick them up to describe images containing the scenes or objects. Consequently, it creates strong connection between the user-specified tags and the images demonstrating the aforementioned scenes and objects, making these tags highly visual-representative. Further, the frequency percentile values associated with the tags suggest that a large user group indeed develops consensus implicitly to use a relatively small set of common tags to describe a large number of images. Specifically, 20 most representative tags have frequency percentile around 99%, indicating that these are extremely popular tags. Also many other representative tags have frequency percentile above 90%.

Observe that people is considered as a least representative tag by our proposed technique. This is surprising as at first glance it may seem that people has a well-defined semantic and typically should represent images containing people. To explore further, we queried Flickr using the most representative and least representative tags (sunset and people) as search keywords, respectively. We observed that the images returned in response to the least representative tag are of great variety especially the background settings. However, most images returned for the tag search sunset indeed show a sunset scene. The images returned in the first pages of Flickr search (conducted during the writing) are depicted in Figures 4(a) and 4(b) for the tags sunset and people, respectively.

An interesting observation on the least visual-representative tags is that most of these tags are locations (e.g., asia, washington, japan, france, china), or time-related such as 2008, july, august, may, or high-level descriptions including picture, photograph, colorful, pic, or camera brands such as finepix, panasonic, and lumix. All these tags do not convey much information related to the visual content of the images. For instance, images accompanied with the asia tag are very diverse and can range from the busy street scenes in Bangkok to images of Gobi desert in Mongolia. Interestingly, most of the least representative tags are also frequently used tags with frequency percentile above 80 or even 90. In summary, the above results demonstrate that the proposed NITC seems to be a good measure reflecting the semantic relationship of an assigned tag to the visual content of the image.

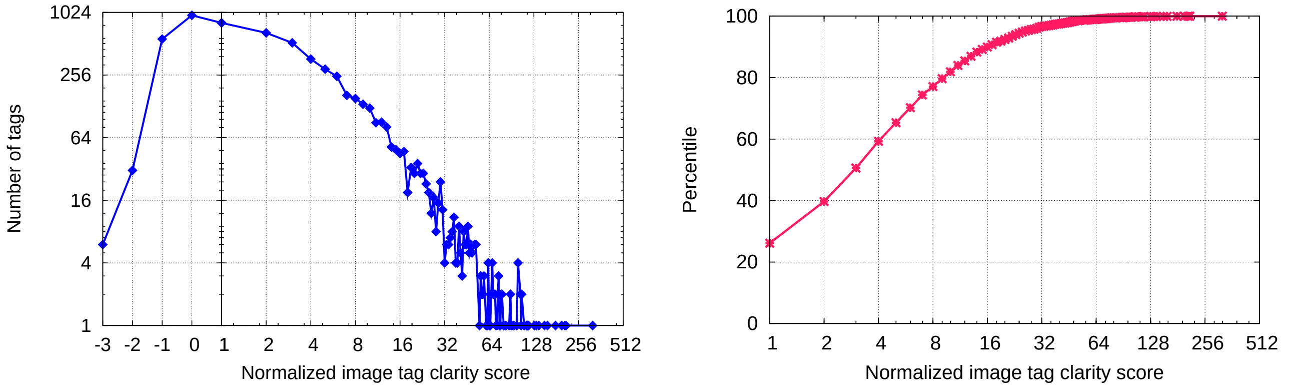
Recall that the images in the dataset are manually classified into 81 pre-defined categories. Interestingly, each category label does match a tag used to annotate some images. Among the top-50 most representative tags shown in Table 1, 12 tags match the category labels in the NUS-WIDE dataset. These 12 tags are shown in bold and prefixed with a star.

As the category labels are believed to be representative in describing the images belonging to the category, we consider the 81 category labels as representative tags and observe their NITC score distribution (Figure 5). Among the 81 tags, 46 are highly representative with NITC score greater than 10; 26 are representative with NITC scores within the range of 2 to 10; and 9 are identified as non-representative. In summary, 89% of the category labels are representative (or highly representative). The only 11% non-representative tags are: dancing, running, soccer, sports, earthquake, castle, town, house and temple. The first five belongs to *Events* and *Activities* and the remaining four belongs to *Scene* and

[4] The number reported here is slightly different from that reported in [1] probably due to different pre-processing. Nevertheless, the tag distribution remains similar.

Table 1: The top-50 most and least visual-representative tags with their NITC scores, frequencies and frequency percentiles. The 12 tags matching category labels in the NUS-WIDE dataset are shown in bold and prefixed with a ★.

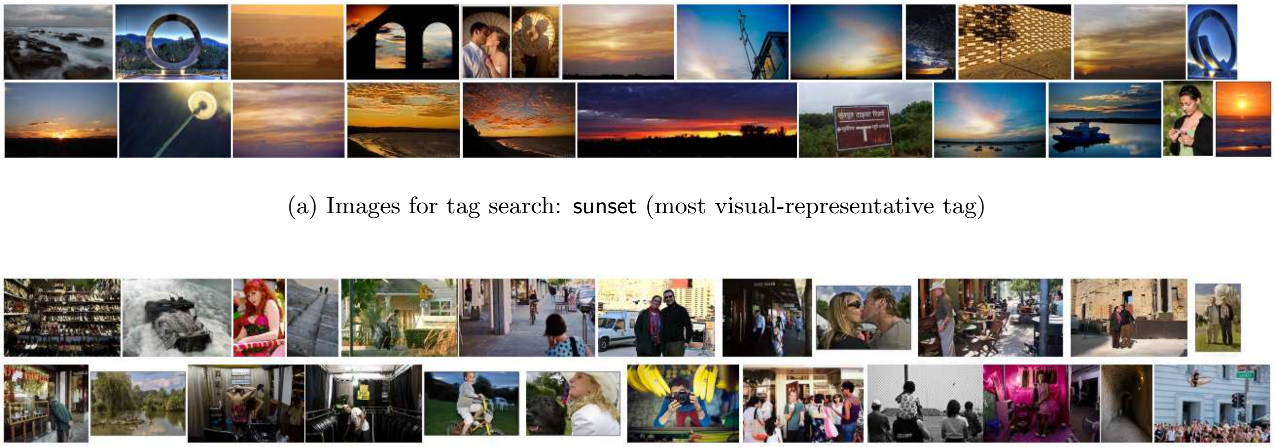
Rank	Most visual-rep tag	NITC(t)	$Freq(t)$	$PerF(t)$	Least visual-rep tag	NITC(t)	$Freq(t)$	$PerF(t)$
1	★ **sunset**	319.6	10,962	99.67	people	-2.9	9,324	99.55
2	silhouette	211.2	3,105	97.64	asia	-2.5	3,878	98.26
3	fog	207.5	2,430	96.71	brown	-2.4	2,501	96.87
4	★ **sky**	197.6	18,935	99.97	japan	-2.3	3,647	98.11
5	sunrise	179.2	3,278	97.76	washington	-2.2	2,605	97.06
6	charts	158.1	557	78.36	2008	-2.1	4,388	98.51
7	★ **sun**	151.9	6,316	99.10	france	-2.0	4,112	98.39
8	mist	138.6	1,632	94.75	picture	-1.7	1,198	92.49
9	★ **clouds**	133.9	14,201	99.85	photograph	-1.6	899	88.86
10	lightning	129.5	482	73.72	july	-1.6	763	86.42
11	blue	118.1	17,822	99.95	china	-1.6	2,562	96.99
12	sea	116.3	9,016	99.52	virginia	-1.5	781	86.99
13	minimalism	114.9	543	77.66	india	-1.5	2,938	97.44
14	landscape	110.2	11,667	99.77	ohio	-1.3	802	87.33
15	windmills	106.7	512	75.76	maryland	-1.3	669	84.17
16	storm	106.0	2,187	96.22	colorful	-1.3	3,001	97.53
17	horizon	105.5	1,195	92.44	pic	-1.3	281	58.70
18	minimal	104.6	533	77.08	august	-1.3	584	80.12
19	★ **beach**	104.3	8,092	99.41	photographers	-1.3	732	85.74
20	dunes	100.9	630	82.59	finepix	-1.3	345	65.36
21	dawn	100.5	1,059	91.09	religion	-1.2	1,608	94.67
22	★ **ocean**	100.2	5,941	99.03	photos	-1.2	1,508	94.22
23	★ **moon**	100.0	1,689	94.87	smorgasbord	-1.2	304	61.33
24	★ **lake**	98.9	4,336	98.50	panasonic	-1.2	709	85.32
25	night	94.1	8,806	99.50	global	-1.2	350	65.74
26	graphs	94.0	107	6.89	may	-1.1	651	83.51
27	graph	91.3	101	1.97	israel	-1.1	780	86.94
28	longexposure	91.0	3,196	97.71	outside	-1.1	1,247	92.81
29	★ **zebra**	89.8	627	82.46	cool	-1.1	1,997	95.70
30	chart	89.6	129	20.70	culture	-1.1	1,297	93.13
31	sketches	87.9	605	81.52	royal	-1.1	463	72.71
32	★ **plane**	83.8	2,014	95.79	world	-1.1	1,822	95.34
33	aircraft	82.4	2,183	96.20	2005	-1.1	2,134	96.05
34	seascape	80.6	1,121	91.72	iranian	-1.0	271	57.33
35	airplane	78.7	2,434	96.72	june	-1.0	768	86.52
36	★ **sand**	78.4	3,595	98.08	pics	-1.0	276	58.08
37	cloud	77.5	3,044	97.61	bottle	-1.0	259	55.63
38	foggy	77.1	383	68.18	april	-1.0	682	84.53
39	weather	76.5	1,907	95.47	september	-1.0	510	75.66
40	morning	75.7	2,403	96.64	hungary	-1.0	317	62.82
41	pattern	74.2	1,209	92.63	caribou	-1.0	596	80.77
42	atardecer	74.1	447	71.96	cannon	-1.0	277	58.23
43	jet	74.1	1,397	93.56	or	-1.0	136	24.08
44	lines	73.7	1,698	94.90	exotic	-1.0	312	62.18
45	dusk	73.4	1,784	95.13	lumix	-1.0	769	86.57
46	moleskine	72.8	426	70.76	republic	-1.0	173	37.07
47	southcascades	71.5	106	6.02	canadian	-0.9	337	64.62
48	★ **water**	70.4	17,646	99.93	this	-0.9	189	41.87
49	unbuilding	70.0	369	67.31	prayer	-0.9	730	85.72
50	craterlakenationalpark	69.4	112	10.58	persian	-0.9	329	64.04



(a) NITC score distribution

(b) NITC percentile

Figure 3: Normalized image tag score distribution.



(a) Images for tag search: sunset (most visual-representative tag)

(b) Images for tag search: people (least visual-representative tag)

Figure 4: Images returned from Flickr by tag search "sunset" and "people".

Location in the concept taxonomy of NUS-WIDE dataset [1]. Compared to the highly representative tags like sunset or clouds, each of the nine non-representative category labels identified lacks clarity in expressing the visual content of images annotated by the tag.

In summary, the above results demonstrate that the proposed normalized image tag clarity score is a good measure for the degree of visual-representativeness of tags. Nevertheless, a formal evaluation remains necessary and is earmarked as future work. Next, we empirically study the relationship between tag frequency and tag clarity.

4.3 Tag Clarity vs Tag Frequency

It is often assumed that extremely popular tags, like stopwords in textual documents, contain little information in image tagging [14]. However, as demonstrated in our empirical study, many of the highly representative tags (e.g., the top-50 most representative tags) have frequency percentile above 99. One example is sky which is the third most popular tag in the dataset. It is also the fourth most representative tag and been used as a category label. Using the notion of image tag clarity, we aim to have a deeper understanding on the relationship between tag clarity and its frequency.

The Pearson's correlations coefficient between tag frequency and tag clarity score is 0.349. That is, they are weakly correlated and more frequent tags are in general more visually representative. In Figure 6, the 5981 popular tags of interests are partitioned into 10 bins according to their frequency percentile, from least popular (1st-9th percentile) to most popular (90th-100th percentile). Each bin contains nearly 600 tags and the percentage of non-representative, representative, and highly representative tags are derived and plotted in Figure 6. It shows that before the tag frequency reaches its 50th percentile, the more popular tags in general become noisier. Slightly more than half of the tags are representative with about 10% being highly representative. Beyond the 50th percentile of tag frequency, we

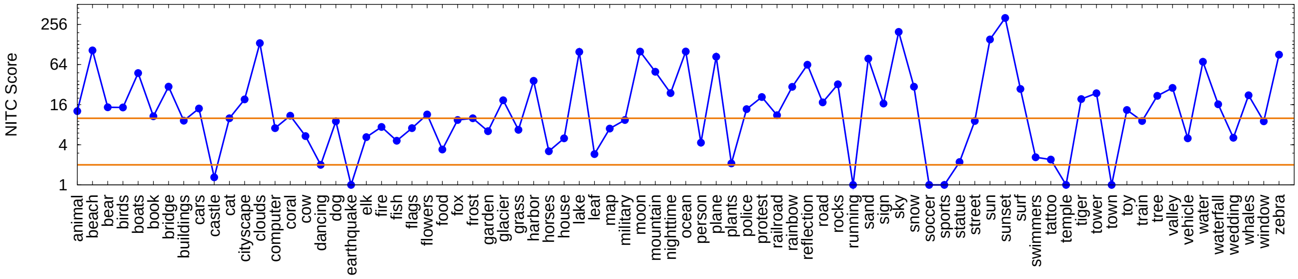


Figure 5: NITC scores of the 81 category labels as tags.

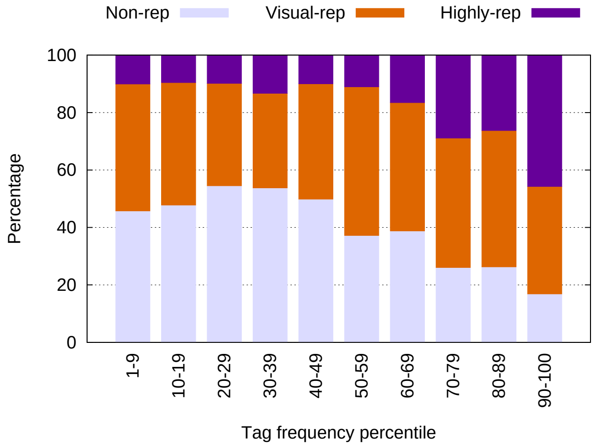


Figure 6: Tag frequency vs tag clarity score.

observe an increase in the percentages of both representative and highly representative tags. Above 90th percentile, nearly half of the popular tags are highly representative and fewer than 16.9% of the tags are non-representative. This is also consistent with the most representative tags listed in Table 1 where many tags have frequency percentile above 99. In summary, highly popular tags are also visually representative.

5. CONCLUSIONS AND FUTURE WORK

With the advent of social media sharing web sites like Flickr, tagging has become an important way for users to succinctly describe content of images. However, the huge volume of tags contributed by ordinary users can be imprecise and hence it may significantly restrict the applications of tags. In this paper, we propose the notion of normalized image tag clarity score to measure the effectiveness of a tag in describing the visual content of its annotated images. Our experimental results demonstrate that most of popular tags are indeed visually representative in describing their annotated images. To the best of our knowledge, this is the first study in identifying visually representative image tags. Our proposed tag clarity score can be effectively used to further improve several tag-based applications. For example, it can be used to rank tags associated with images according to their visual representativeness as well as recommending tags. In the future, we wish to investigate in detail how tag clarity impacts these applications.

6. ACKNOWLEDGEMENT

This work was partially supported by A*STAR Public Sector R&D, Singapore, Project Number 062 101 0031.

7. REFERENCES

[1] 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. of ACM CIVR'09*, Santorini, Greece, July 2009.

[2] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proc. of SIGIR'02*, pages 299–306, Tampere, Finland, 2002.

[3] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *Proc. of SIGIR'08*, pages 347–354, Singapore, 2008.

[4] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[5] C. Hauff, V. Murdock, and R. Baeza-Yates. Improved query difficulty prediction for the web. In *Proc. of CIKM'08*, pages 439–448, Napa Valley, CA, 2008.

[6] 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.

[7] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proc. WWW'09*, pages 351–360, Madrid, Spain, 2009.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

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

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

[11] J. Teevan, S. T. Dumais, and D. J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proc. of SIGIR'08*, pages 163–170, Singapore, 2008.

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

[13] 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.

[14] L. Wu, L. Yang, N. Yu, and X.-S. Hua. Learning to tag. In *Proc. WWW'09*, pages 361–370, Madrid, Spain, 2009.

[15] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *Proc. of SIGIR'05*, pages 512–519, Salvador, Brazil, 2005.

[16] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *Proc. of SIGIR'07*, pages 543–550, Amsterdam, 2007.